Generalisable FPCA-based Models for Predicting Peak Power in Vertical Jumping using Accelerometer Data

Mark George Eric White

Submitted to Swansea University in fulfilment of the requirements for the degree of Doctor of Philosophy

Swansea University

2021

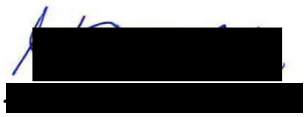
ABSTRACT

Peak power in the countermovement jump is correlated with various measures of sports performance and can be used to monitor athlete training. The gold standard method for determining peak power uses force platforms, but they are unsuitable for field-based testing favoured by practitioners. Alternatives include predicting peak power from jump flight times, or using Newtonian methods based on body-worn inertial sensor data, but so far neither has yielded sufficiently accurate estimates. This thesis aims to develop a generalisable model for predicting peak power based on Functional Principal Component Analysis applied to body-worn accelerometer data. Data was collected from 69 male and female adults, engaged in sports at recreational, club or national levels. They performed up to 16 countermovement jumps each, with and without arm swing, 696 jumps in total. Peak power criterion measures were obtained from force platforms, and characteristic features from accelerometer data were extracted from four sensors attached to the lower back, upper back and both shanks. The best machine learning algorithm, jump type and sensor anatomical location were determined in this context. The investigation considered signal representation (resultant, triaxial or a suitable transform), preprocessing (smoothing, time window and curve registration), feature selection and data augmentation (signal rotations and SMOTER). A novel procedure optimised the model parameters based on Particle Swarm applied to a surrogate Gaussian Process model. Model selection and evaluation were based on nested cross validation (Monte Carlo design). The final optimal model had an RMSE of 2.5 $W \cdot kg^{-1}$, which compares favourably to earlier research (4.9 ± 1.7 $W \cdot kg^{-1}$ for flight-time formulae and 10.7 ± 6.3 $W \cdot kg^{-1}$ for Newtonian sensor-based methods). Whilst this is not yet sufficiently accurate for applied practice, this thesis has developed and comprehensively evaluated new techniques, which will be valuable to future biomechanical applications.

DECLARATIONS AND STATEMENTS

Declaration

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed .....  ..... (candidate)

Date .....14th April, 2021.....

Statement 1

This thesis is the result of my own investigations, except where otherwise stated.

Other sources are acknowledged giving explicit references. A bibliography is appended.

Signed .....  ..... (candidate)

Date .....14th April, 2021.....

Statement 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed .....  ..... (candidate)

Date .....14th April, 2021.....

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would like to express my gratitude and appreciation to my principal supervisor, Dr Neil Bezodis, who has conscientiously supported and guided my research. He has shown wisdom and common sense in helping shape my ideas with always an eye to the overarching narrative. His thoughtful suggestions were invaluable. I appreciate the academic freedom he has given me to explore these many machine learning models. I hope some of his undoubted ability and skill has rubbed off on me.

I am grateful to Dr Jonathan Neville, my second supervisor from the Auckland Institute of Technology (AUT). He brought a fruitful combination of machine learning know-how and practical experience of using wearable sensors in biomechanics. I could be sure he would get to the nub of what I was doing. He would pose the most challenging questions but offer helpful suggestions. I am sure this thesis is the better for it.

Machine learning is a vast and ever-growing discipline. I am thankful to Prof. Paul Rees, my third supervisor, who has generously shared his expertise in this exciting discipline. He has given me encouragement as I developed new techniques, providing reassurance I was on the right track. I also wish to thank Dr Claire Barnes who contributed her knowledge of machine learning, applied to wearable sensors.

I would like to offer my sincere thanks to Profs. Huw Summers and Liam Kilduff, who were instrumental in establishing the collaboration between Swansea University and the AUT. The joint venture funded my research. As my first supervisor, Prof Summers set the ball rolling and provided a wider engineering perspective. Despite his growing administrative responsibilities, he made sure I was well served by a strong supervisory team.

The ideas of other academics are always prized. I very much appreciated my long conversations with Dr Nick Owen about many aspects of biomechanics, not least the value of neuromuscular peak power. We saw the world in much the same way, perhaps because we both first trained as physicists.

Scientific research is impossible without the help and cooperation of laboratory technicians. In this and all respects, I could not have been better served by Wendy Clark and her team.

A PhD research programme is a considerable undertaking. My friends and family have supported me all the way in good times and bad. The support from my fellow students, sharing the same endeavour, has also been invaluable. In particular, I would like to thank my closest friends, Helen Parrott and Dr Louise Burnie, for their support and good humour in making the journey more enjoyable. It would not have started without the encouragement of my friends at Herne Hill Harriers. To Stephanie Mitchell, my former training partner who first suggested the idea, and to the rest of my training group, I extend my warmest gratitude.

CONFERENCE PAPERS

White, M. G. E., Bezodis, N. E. (2018). Validation of using Inertial Measurement Units for measuring performance in vertical and horizontal jumps. *Pan-Wales Sports Science Conference*, Bangor University: Vol. 2.

White, M. G. E., Bezodis, N. E. (2019). The contribution of arm swing to vertical jump performance. *Pan-Wales Sports Science Conference*, Cardiff Metropolitan University: Vol. 3.

White, M. G. E., Bezodis, N. E., Neville, J., Summers, H. (2020). Force-time curve alignment for Functional Principal Component Analysis in Vertical Jumping. *ISBS Proceedings Archive*: Vol. 38 Available at: https://commons.nmu.edu/isbs/vol38/iss1/82

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Acronym	Definition
ACC*n*	Resultant acceleration component corresponding the designated FPC.
AD1*n*	Resultant acceleration first derivative corresponding the designated FPC.
AD2*n*	Resultant acceleration second derivative corresponding the designated FPC.
AM	Accelerometer Model – function carrying out all operations on accelerometer data and then running cross validation using a specified algorithm to estimate predictive error.
ANOVA	Analysis of Variance
ARD	Automatic Relevance Determination – anisotropic kernel used in GP
BC	Box Constraint – SVM hyperparameter
CL*n*	Cluster identifier
CM	Centre of Mass
CMJ_A	Countermovement Jump with arm swing
CMJ_{NA}	Countermovement Jump with no arm swing
CV	Cross Validation – procedure for determining a models' generalised error.
DIS	Pseudo displacement component from the twice-integrated resultant acceleration.
DNN	Deep Neural Network
FDA	Functional Data Analysis – branch of mathematics concerned with representing time series data as smooth basis functions.
FPCA	Functional Principal Component Analysis – procedure for extracts FPCs (features) and FPC scores from a set of smoothed functions.
FPC	Functional Principal Component – time-dependent function describing the variance in a set of smoothed functions.
GCV	Generalised Cross Validation – procedure used in this thesis for fitting smooth basis functions to raw data.
GLM	Generalised Linear Model – a flexible model with mild assumptions used in this thesis to compare model predictive errors.
GP	Gaussian Process – probability distribution of best fit over a range of values according to Bayesian principles.
GPR	Gaussian Process Regression – general model type
HPO	Hyperparameter Optimisation – a limited Matlab procedure using Bayesian optimisation to determine optimal hyperparameter values.
KS	Kernel Scale – SVM hyperparameter
LB	Lower Back – anatomical position (L4) of a sensor
LR	Linear Regression – general model type
LS	Left Shank – anatomical position of a sensor
LSTM	Long/Short Term Model – type of deep neural network
M32	Matérn 3/2 – GP kernel function
M52	Matérn 5/2 – GP kernel function
MAE	Mean Absolute Error

Acronym	Definition
MCCV	Monte Carlo Cross Validation – procedure based on many repeated splits of the data to obtain an estimate of the model's predictive error.
NC	Non-Clustered – type of bagging method used in this thesis.
NCA	Non-Clustered identifier of all the models in SM series
NC5	Non-Clustered identifier of the last 5 models in SM series
NCV	Nested Cross Validation – procedure separating model selection from model evaluation (estimating generalised error).
NN	Neural Network
PSO	Particle Swarm Optimisation – procedure used in this thesis to find the global optimum of the SM.
PWR*n*	Pseudo power functional component – product of acceleration and pseudo velocity components.
RFD	Rate of Force Development
RMSE	Root Mean Squared Error
RQ	Rational Quadratic – GP kernel function
RS	Right Shank – anatomical position of a sensor
SBC	Schwarz Bayesian Information Criterion for predictor selection
SD	Standard Deviation
SE	Standard Error
SJ	Squat Jump
SM	Surrogate Model – function that predicts the loss of another model, which in this thesis is the accelerometer model.
SMOTE	Synthetic Minority Oversampling Technique – data augmentation method for classifiers
SMOTER	SMOTE for Regression models
SPCA	Supervised Principal Component Analysis
SqExp	Squared Exponential – GP kernel function
SR	Signal Rotation – data augmentation method introduced in this thesis
Std	Standardisation –conversion of model predictors and outcome variables to Z-scores prior to fitting the model.
SVM	Support Vector Machine – general model type
UB	Upper Back – anatomical position of a sensor
VEL*n*	Pseudo velocity functional component from the integrated resultant acceleration.
VGRF	Vertical Ground Reaction Force

Mathematical terms

Term	Definition
$\mathcal{A}$	Algorithm
$\boldsymbol{\beta}$	Parameters
$\mathcal{D}_0$	Full accelerometer data set excluding holdout
$\mathcal{D}_\mathrm{H}$	Holdout accelerometer data set
δ	Standard deviation for upper limit to random search
ε	Margin boundary parameter in SVM models
i	Random search iteration counter
I	Number of random search observations
j	Search iteration counter between SM re-training
k	Outer fold index
K	Number of outer folds
l	Inner fold index
L	Number of inner folds
$\mathcal{L}$	Cross-validated model loss
λ	Regularisation parameter in LR models
λ_FS	Regularisation parameter for functional smoothing
m	Model index referring to SM optimisation series
M	Number of models in the SM series
n	Jump index
N	Number of jumps
σ	Noise parameter in GPR models
t_pre	Time before the event of interest, usually take-off – used to define the time window.
t_post	Time after the event of interest, usually take-off – used to define the time window.
Ψ	Set of discretised wavefunctions defining the FPCs
$\mathbf{X}$	Set of FPC scores
$\mathbf{Z}$	Set of random search observations

Mathematical notation

Notation	Definition
$A^{(k)}$	Bracketed superscript indicates quantity pertains to the k-th outer fold.
$A^{(k,l)}$	Bracketed superscript with comma-separate terms indicates quantity pertains to l-th inner fold within the k-th outer training set.
$\hat{A}$	A hat symbol above the quantity indicates the optimal value.
A_{+}	Plus subscript indicates the training set
A_{-}	Minus subscript indicates the validation set
A_{i}	Letter subscript refers to an index, usually as part of a series.

“Nature uses only the longest threads to weave her patterns, so each small piece of her fabric reveals the organisation of the entire tapestry.”

— Richard P. Feynman, 1964.

“The only real validation of a statistical analysis, or of any scientific enquiry, is confirmation by independent observations.”

— Francis J. Anscombe, 1967.

CHAPTER 1. INTRODUCTION

1.1 Context

1.1.1 Neuromuscular power

The ability to generate high levels of neuromuscular power is a critical aspect of performance in many sports. Peak neuromuscular power is strongly correlated with sprint acceleration which is an important aspect of many team sports (Baker, 2002; Baker & Nance, 1999a, 1999b; Chelly & Denis, 2001; Young et al., 2005). Peak power output also serves as a useful indicator for signs of overtraining and fatigue and providing evidence of the efficacy or otherwise of a training programme (Cormack, Newton, & McGuigan, 2008; Gathercole et al., 2015; McLean et al., 2010; Twist & Highton, 2013). Consequently, an athlete's peak neuromuscular power is monitored regularly in many elite and professional sporting environments to track progress in training, increase performance readiness, and help reduce injury and illness (T. Jones et al., 2016; Taylor et al., 2012).

1.1.2 Traditional methods

Many testing methods are based on a standing vertical jump, which has long been established as the preferred indicator of maximal neuromuscular power (Claudino et al., 2017). The jump can be quickly performed without residual fatigue, allowing jump testing to be performed typically at the start of a training session (J. McMahon et al., 2019). The countermovement jump is a popular choice, although some coaches prefer a mix of jumps depending on the needs of the sport (T. Jones et al., 2016; Taylor et al., 2012). Most coaches prefer jump height as their metric since it can be measured with minimal equipment, either based on flight time or, if arm swing is permitted, using a jump-and-reach test (T. Jones et al., 2016). Both methods are well-known to have their limitations (Hatze, 1998; G. Markovic et al., 2004), but the simple test of jump height is convenient for field-based testing, which may well be the reason why it remains popular.

Jump height is a useful proxy for peak power, but it is not the same as the criterion measure (G. Markovic et al., 2004). The gold standard method uses a force platform based on twice integrating the time-dependent vertical ground reaction forces (VGRF). However, the equipment is expensive and cumbersome and requires expertise to operate. Testing has to be confined to the laboratory or some other suitable facility, making it unsuitable for field-based use. The only other alternative for many years was to use a simple peak power prediction equation based on jump height and body mass (e.g. Amonette et al., 2012; Canavan & Vescovi, 2004; Sayers et al., 1999). Many such equations were proposed, but the same equation's predictive error could vary considerably between different groups (e.g. Tessier et al., 2013).

1.1.3 Methods based on inertial sensors

In recent years, researchers have attempted to use inertial sensors worn on the body to estimate peak power output (Choukou et al., 2014; Giroux et al., 2014; Hojka et al., 2018; Mauch et al., 2014). The estimates were based on using the sensor data to compute the body's vertical acceleration and velocity, allowing the instantaneous power to be computed directly by factoring in the body mass. These Newtonian methods depended critically on knowing the sensor's changing orientation in the global reference frame so the true vertical direction could be determined (Picerno, Cereatti, et al., 2011). While orientation-correcting algorithms have been developed with increasing sophistication, they lacked the necessary precision for this purpose. Consequently, the peak power estimates based on these methods were less accurate than the simple peak power equations above.

Inertial sensors have received considerable attention in biomechanics over the last decade, finding applications in many situations, often in conjunction with a machine learning (ML) model (Claudino et al., 2017; Halilaj et al., 2018). Typically, a suitable method identifies certain features in the data that are used as inputs to the model or neural network. The aim is to identify which of the measures do have relevance to the outcome variable in question. The approach has been used widely in activity recognition (classification models, e.g. Attal et al., 2015; Burdack et al., 2019; Groh et al., 2017; McGrath et al., 2019; O'Reilly et al., 2017), but only to a limited extent

in predicting performance measures (regression models, e.g. Derie et al., 2020). However, when this approach was applied in the past to the gold-standard VGRF data, the models based on kinematic, kinetic or coordinative variables could only explain 66–88% of jump height variance (Aragón-Vargas & Gross, 1997; Dowling & Vamos, 1993; Oddsson, 1987). In light of this, it would be worth considering another method of extracting features from sensor data. Considering the practicalities of field-based testing, it would be preferable if the data came from a single sensor so that the jump test can be administered quickly with minimal setup time. It is worth noting that in elite team sports, players often wear a single inertial measurement unit (IMU) that is typically used for GPS tracking, but the unit may also include other sensors.

1.1.4 Principal Component Analysis

Principal Component Analsyis (PCA) is a popular technique for reducing the high dimensionality of large data sets, increasing their interpretability whilst minimising information loss (Jolliffe & Cadima, 2016). PCA involves projecting the data onto a lower dimensional orthogonal subspace where the variances are maximised, an interpretation first proposed by Hotelling in 1933 (Bishop, 2006). The subspace is defined by eigenvectors of the covariance matrix called principal components (PC). The projected data are the eigenvalues, usually called the PC scores, which are often referred to as features of the data in a machine learning context. The PC scores are often used as inputs in subsequent ML models (Bishop, 2006; Hastie et al., 2009). The original data may be reconstructed from a linear combination of these features, where the level of accuracy depends on the number of components used. Indeed, the features can often be interpreted in a meaningful way, especially after applying a suitable subspace rotation (Hastie et al., 2009). For example, Markovic et al. (2004) interpreted the first PC of their analysis of jump performance data as ‘explosiveness’. Hence, the participants could be rated for their explosiveness by using this component. In this respect, PCA has similarities with Factor Analysis, although they are fundamentally different techniques (Jolliffe, 2002).

There are various adaptations of PCA arising from its widespread use across many disciplines. The biomechanics field has embraced Functional Principal Component

Analysis (FPCA) (Harrison, 2014; Ramsay & Silverman, 2005). FPCA identifies characteristic patterns in time series data called Functional Principal Components (FPCs). They describe modes of variation across a set of time series or curves, which when plotted, have an intuitive appeal. As with PCA, each functional component is orthogonal to the others, making them distinct and independent. Across a diverse range of applications, studies have reported that these types of features can be strongly associated with performance measures or injury risk in rowing, swimming, weightlifting, race walking and jumping (Coffey et al., 2011; Donà et al., 2009; Donoghue et al., 2008; Floria et al., 2014; Harrison et al., 2007; Kipp et al., 2012b, 2012a; Kipp & Harris, 2015; Liebl et al., 2014; Mallor et al., 2010; Ryan et al., 2006; Sacilotto et al., 2015; Warmenhoven et al., 2017a, 2017b, 2017c). The curves described the time evolution of either VGRF data, kinematic or kinetic variables in these applications.

It would appear that FPCA would be well-suited to sensor data but there is no previous instance of FPCA being used in this way. Moreover, it is notable that aside from simple regression models (Moudy et al., 2018; Richter et al., 2014a), FPCA has not been used as a feature extraction method for more sophisticated models that have become commonplace in the ML literature. In contrast, non-functional PCA is often the standard method for dimensional reduction. However, it is less well suited to time series analysis because its component scores would themselves become time-dependent. For a time series, this means many predictors for the model, one for each sampling interval, which is usually handled more appropriately by a neural network. In FPCA, on the other hand, there are few component scores because each principal component fully describes the time evolution of the feature. FPCA is a more concise representation, offering an intuitive understanding gained from the patterns described by the FPCs. Therefore, it would seem reasonable to investigate using FPCA as the feature extraction technique for an ML model trained to predict peak power output from sensor data. It is an approach that has not been previously investigated, but it would appear to have promise.

1.1.5 Machine learning

ML techniques have developed considerably over the last four decades thanks to advances in statistics and computer science. Supporting these developments has been an exponential growth in computing power with a commensurate reduction in cost. The so-called classical ML models depend on prior feature extraction, as alluded to above, while deep neural networks (DNNs) learn to recognise features themselves (Bishop, 2006; Hastie et al., 2009). Many applications are in activity recognition concerned with tracking everyday tasks (e.g. Hammerla et al., 2016; Ordóñez & Roggen, 2016; J. Wang et al., 2019), but there are examples of systems to recognise sports movements such as beach volleyball serve types, tennis strokes and snowboarding tricks (Cuspinera et al., 2016; Kautz et al., 2017; D. Yang et al., 2017).

There is an opportunity to use ML for predicting performance measures more extensively in biomechanics. Vertical jumping would seem to be a suitable exemplar, as the CMJ is a controlled movement that has been studied extensively. Since FPCA has shown promise as a feature extraction technique, it would be sensible to take the classical ML route rather than to use deep neural networks. DNNs typically require substantial volumes of training data (Hastie et al., 2009), which does not lend itself to biomechanics, where it can be challenging to obtain large data volumes. The computational costs of training classical models are comparatively light, making it possible to apply more sophisticated validation methods.

Cross validation is a fundamental part of selecting and appraising a model where the aim is to find a suitable model that performs well on previously unseen data drawn from the same distribution (Arlot & Celisse, 2010; Y. Zhang & Yang, 2015). It is also vital to have an unbiased estimate of the model's generalised predictive error and so avoid making overly optimistic claims of its capabilities (Cawley & Talbot, 2010). Cross validation makes efficient use of the available data, repeatedly splitting the data at random into training and validation sets and then averaging the results. The K-fold design has become commonplace in machine learning for its simplicity and relatively low computational demands (Arlot & Celisse, 2010; Kohavi, 1995). However, by reusing the data, model selection bias becomes an issue leading to over-optimistic error estimates, which sometimes can be substantial (Cawley & Talbot, 2010; Krstajic et al.,

2014; Vabalas et al., 2019). This situation can be avoided by using nested cross validation, which provides the structure to allow a series of independent validation checks (Stone, 1974; Varma & Simon, 2006). It comes at a substantially increased computational cost, but it yields unbiased estimates (Vabalas et al., 2019). Nested cross validation has scarcely been used before in biomechanics, but in light of a wide range of errors reported for the same peak power prediction equations (Section 1.1.2), it would be appropriate to take a more robust approach in this thesis.

1.2 Statement of purpose

The aim of this thesis is *to develop a machine learning approach for predicting peak power in the countermovement jump based on accelerometer signals from a single body-worn sensor*.

1.3 Research questions

1.3.1 Concerning accuracy

Countermovement jumps without arm swing are often used to determine peak power output, partly because the jump is more controllable and hence reliable, and because it places emphasis on the lower extremity. However, for some practitioners, a countermovement jump with arm swing is a more natural athletic movement that is more relevant to the sport, such as volleyball. Therefore, it will be appropriate to assess how well the models perform for the CMJ, with and without arm swing.

FPCA has been applied to models predicting CMJ jump height using VGRF data (Moudy et al., 2018; Richter et al., 2014a), but the same approach has not been applied to estimating peak power. Therefore, evaluating an FPCA-based model based on gold-standard data would support the case for the FPCA approach before proceeding to use accelerometer data. Hence, the first research question is: *how well does an FPCA-type model perform when predicting peak power and jump height in the CMJ, with and without arm swing, based on gold-standard VGRF data?* Two performance measures are included to allow direct comparison with previous studies.

Assuming the FPCA-model performs well, such models can be developed for predicting peak power using data from accelerometers. The data will depend on where the sensor is attached to the body and the type of jump. The inertial accelerations recorded by sensors at different anatomical locations will have different characteristics. Hence, one sensor's set of characteristics may be more conducive than others for predicting peak power output. Therefore, the second research question is, *how does an FPCA-type model perform when predicting peak power in the CMJ, with and without arm swing, based on accelerometer data from sensors at different anatomical positions?* This is an overarching question, as the answers will depend on the factors raised below.

1.3.2 Concerning model function specification

The term 'model function' is defined to encompass all the procedures involved in the modelling process, including data augmentation, data preprocessing, feature extraction, feature selection and model optimisation and fitting. It reflects the ideas of nested cross validation, which determines the generalised predictive error of the whole modelling procedure (Cawley & Talbot, 2010; Krstajic et al., 2014; Varma & Simon, 2006).

The model's performance depends on the underlying algorithm, but there are no examples of FPCA-type models based on body-worn accelerometer data to provide a guide. Therefore, in keeping with standard practice, a wide variety of possible algorithms will need to be evaluated. Therefore, the third question is, *which machine learning algorithms are better suited to predicting peak power in vertical jumping based on the FPCA characteristics of body-worn accelerometer data?* Answering this question will require each algorithm to be optimised as its performance can vary considerably depending on the values chosen for its hyperparameters.

The preprocessing of the accelerometer data may be expected to influence the model's predictive accuracy. Consideration will need to be given to the time window that demarcates the accelerometer signal. The degree of smoothing is another factor as, depending on its severity, it will preserve or eliminate certain characteristics, which may or may not contain relevant information for the model. Curve registration

procedures may prove valuable as they improve the curve alignment, allowing the curve variance to be decomposed into its amplitude and temporal components (Kneip & Ramsay, 2008; Ramsay & Li, 1998). Thus, these preprocessing procedures have a bearing on the accelerometer characteristics serving as predictors in the model and will influence its accuracy. These considerations are encapsulated in the fourth research question, *how should the accelerometer data be preprocessed to minimise the model's predictive error?*

It will be valuable to understand which characteristics of the accelerometer signal are essential to the model. The features extracted will depend on the representation of the accelerometer signal itself, such as whether to take the resultant or to use the original triaxial data, or employ other transformations. Different representations of the data may bring to prominence certain aspects that could be conducive to the models. These techniques may be considered data preprocessing, but it will be more appropriate to examine them separately as they focus on changing what the features represent rather than simply making small adjustments to them, as above. All of these different approaches to feature extraction are incorporated in the fifth research question, *what characteristics of the accelerometer signal are more important to the model in predicting peak power?*

In many fields of science, it can be difficult to obtain sufficient data to achieve the statistical power needed in hypothesis testing or to develop a model with the required predictive accuracy. The issue is arguably more acute in the human sciences where it can be challenging to recruit enough, suitably qualified volunteers. As a general rule, a model trained on a larger data set will make predictions with smaller errors on average as the training data provides more examples, while the model can afford to have more predictors to refine estimates without leading to overfitting (Bishop, 2006). Therefore, understanding how sensitive the model is to sample size is an important consideration for experimental design. It may be possible to improve accuracy with data augmentation techniques in certain ranges of the outcome variable where there are fewer training examples (Chawla et al., 2002; Torgo et al., 2015; Torgo & Ribeiro, 2007). In terms of jump testing, that may translate to improving the model's accuracy at the higher end of the jump performance scale, where there is likely to be more

interest from elite sport. So the final research question is, *how does sample size and composition affect the model's predictive error, and can data augmentation bring worthwhile improvements in accuracy?*

1.4 Structure

Chapter 2 – Literature Review

The chapter considers current practice in athlete monitoring regarding the use of jump testing, its validity and reliability, and its correlation with sports performance. Alternative methods for predicting peak power are examined based on simple regression equations and earlier attempts using wearable sensors. Methods of predicting other discrete VGRF-related measures are considered, many of which used machine learning, highlighting useful techniques and approaches. The review examines FPCA in more detail, setting out its principles and the definition of the FPCs as they will form the bedrock of the approach taken in this thesis. The chapter concludes by examining the methodological aspects essential to selecting and appraising a model, including cross-validation design, optimisation and enhancements from data augmentation.

Chapter 3 – Data collection

The investigation begins with a full description of the methods undertaken to gather the VGRF and accelerometer data that will be used throughout the thesis. Details are provided concerning the participants involved, the protocols followed, and the calculations of criterion jump performance. The jump height and peak power measures will allow comparisons with other studies, while peak power will serve as the 'ground truth' for the accelerometer models. Details on the selection and calibration of the sensors are presented in Appendices A and B.

Chapter 4 – VGRF models of jump performance

This chapter assesses the FPCA technique by extracting the features from the gold-standard VGRF data. The key features are examined to understand how they contribute

to jump performance using the intuitive nature of FPCs rather than simply treating them as predictor variables. This also provides an opportunity to evaluate the effectiveness of using curve registration to improve model accuracy.

Chapter 5 – Accelerometer model selection and optimisation

The development of an accelerometer model begins with this chapter. It sets out a novel and comprehensive optimisation procedure for model selection and appraisal, based on cross validation that was developed to ensure robust and reliable results. The optimisation procedure will be used throughout the subsequent chapters. The analysis starts by identifying the combination of sensor attachment site and jump type that yields a more accurate model. As a result, only the data set with the most favourable combination of sensor and type will be taken forward, streamlining the investigation. The chapter also draws up a shortlist of ML models, tuning their hyperparameters to minimise prediction error using the new optimisation procedure. Over the following two chapters, this shortlist is whittled down to a final preferred algorithm.

Chapter 6 – Data preprocessing

This chapter seeks to improve the shortlisted models by optimising the preprocessing of the accelerometer signal to reduce the predictive error. It establishes the best time window demarcating the start and end of the time series with respect to either take-off or landing. New algorithms are introduced to detect these two events based on the accelerometer signal to free the model from any remaining dependency on the VGRF data, an essential step for a field-based system. The chapter also determines the appropriate level of smoothing to be applied to the signal and evaluates whether curve registration can reduce errors further.

Chapter 7 – Feature selection

The final chapter on accelerometer model development investigates whether FPCA applied to different data representations can improve accuracy, especially when the FPCs are chosen more selectively. Chapters 5 and 6 work with the resultant acceleration taking a set number of retained components as a consecutive block

without considering feature selection. In this chapter, models are evaluated based on the features extracted either from resultant or triaxial data, or from time derivatives and integrations of the curves, including a pseudo power representation. These new representations yield a much larger set of predictors for the models, emphasizing the need for feature selection to avoid overfitting. Two established feature selection methods are employed to compare the different representations. In light of this analysis, a final model is chosen.

Chapter 8 – Sampling size and Data Augmentation

This chapter determines the final model's learning curve, which describes its predictive error as a function of sample size, using a resampling method. The analysis also assesses the benefits of participants performing multiple trials to enlarge the data set. The investigation then considers two data augmentation techniques to artificially increase the size of the data set by increasing the number of points in sparse regions of the peak power range. The first method is an established technique that interpolates between existing observations within the feature space. The second is a new technique that generates new curves with random signal rotations as if the sensor was attached at different angles.

Chapter 9 – General discussion

The thesis concludes with a review of the results obtained, formally answering the research questions posed above. It puts into context what has been achieved in this thesis, highlighting its value to both biomechanics and machine learning communities. It takes a critical view of the methods employed and notes the difficulties of using body-worn sensors to estimate performance measures. An illustration shows how model errors reduce the sensitivity of detecting worthwhile changes in peak power. The prospects for applying these techniques more widely to other biomechanical applications are considered. Directions are recommended for future research to overcome the limitations encountered.

1.5 Code repository

The methods and procedures were implemented in Matlab R2020a (MathWorks, Natick, MA, USA) and SAS Studio 3.8, University Edition (SAS Institute Inc., Cary, NC, USA), as specified in the text. Custom scripts were developed for all the investigations, including the novel optimisation procedure and the new algorithms that were developed. The custom code is available for download from the GitHub repositories found at https://github.com/markgewhite.

CHAPTER 2. LITERATURE REVIEW

2.1 Introduction

It is one hundred years since Sargent (1921) first developed his eponymous jump. He saw it as a measure of 'physical efficiency', but it would later be recognised as a measure of neuromuscular power (Adamson & Whitney, 1971). Since those early developments, the vertical jump has become one of the fundamental movements studied in biomechanics. Development in the field has gone hand in hand with advances in technology, improving the accuracy and reliability of jumping power measurement. For practitioners today, assessing neuromuscular power is often a key consideration in an athlete's physical preparation (J. McMahon et al., 2019).

Research has been dedicated to developing suitable, field-based alternatives to the gold standard force platform method. Early developments created simple formulae for predicting peak power output based on jump height and body mass, but their accuracy was highly variable. Attention shifted towards using wearable sensors based on a Newtonian approach to track the body's jumping movement. However, the technology lacked sufficient precision to allow accurate estimates of peak power to be made. Despite significant developments in using machine learning elsewhere in biomechanics, such advanced methods have yet not been applied to predicting peak power in vertical jumping.

This chapter begins by reviewing the current practice in athlete jump testing across many different sports (Section 2.2). It considers the validity and reliability of peak power and jump height, the two popular measures of jump performance. Those measures have strong associations with sports performance and can be used as indicators of fatigue. Section 2.3 summarises existing methods before examining different approaches that have been taken to estimate peak power, including the use of wearable sensors. Section 2.4 takes a broader view of machine learning in biomechanics, highlighting useful techniques and approaches that have been used in predicting VGRF-related measures. Section 2.5 then looks at functional principal components as they can capture the essential features in time series data, which has

been useful in a diverse range of applications. Section 2.6 considers cross validation for the distinct aims of model selection and model assessment. It also examines optimisation methods and data augmentation strategies to improve predictive accuracy. Section 2.7 summarises the key findings, drawing conclusions on the methods and approach that should be taken to fulfil the research aims of this thesis.

2.2 Jump testing

2.2.1 Current practice

Jump testing is popular amongst coaches as it can be carried out quickly with minimal familiarisation for athletic populations who have acquired the necessary motor skills (Moir et al., 2004). The tests are easy to administer, cause minimal fatigue and have little or no adverse impact on subsequent training sessions (J. McMahon et al., 2019; Twist & Highton, 2013). In practice, there are many different protocols and measures that coaches are interested in. A recent comprehensive survey of elite rugby union reported that the vast majority of practitioners (36 out of 42 coaches and sports scientists who responded) used jumping as their preferred test of a player's neuromuscular power (T. Jones et al., 2016). The most popular jump was the countermovement jump (CMJ), followed by broad jumps, drop jumps, squat jumps or 'triple-response jumps'. More practitioners favoured jump height as their preferred metric (19/36 or 53%), compared to the reactive strength index (RSI) or some quantities associated with specific jump types. An Australian survey across 14 sports reported that all respondents who provided more detailed information (11) used the CMJ with jump height as their chosen metric (Taylor et al., 2012). The CMJ is commonplace in academic research where it featured in thousands of articles (1886 PubMed, 2866 Scopus and 2979 from the Web of Science), according to a comparatively recent systematic review (Claudino et al., 2017). The review, which focused on intervention studies with at least a 3-week duration in either men or women but not both, reported that most studies involved athletes (60%) with a strong emphasis on male participants (80%). Nearly half of the sport-related studies involved soccer players (49%), with other prominent sports including basketball (10%), track and field

(8%), volleyball (5%), handball (5%), judo (3%) and rugby union (3%). In summary, the CMJ is the most popular jump, and hence the jump would appear to be a good choice for this thesis as the findings would be of interest to the largest number of sports. However, the validity and reliability of the CMJ and related jumps should be considered first.

2.2.2 Validity and reliability

Validity and reliability depend on the jump type and the chosen measure of jump performance. One of the most comprehensive studies was undertaken by Markovic et al. (2004), who evaluated seven 'explosive power tests', including five vertical and two horizontal jumps. The CMJ and squat jump (SJ) tests achieved the highest levels of reliability (Cronbach's α = 0.97 and 0.98, respectively, compared 0.93–0.96 for the other jumps) based on jump height. The coefficients of variation (CVs) were also the lowest for these two vertical jumps, 2.8% and 3.3%, respectively. The sample was much larger than in previous investigations, involving 93 male physical education students, making it one of the most comprehensive studies. Other studies that typically involved fewer than 20 participants reported CVs of 4.3% and 5.0% for CMJ and SJ, respectively (Arteaga et al., 2000), 2.4% for both jump types (Moir et al., 2004), and 6.3%, also for both types (Viitasalo, 1985). Myers et al. (1993) found that the CMJ and the broad jump were the tests most strongly associated with 'explosive strength' using factor analysis compared to a wide range of exercise assessments. Markovic et al. (2004) also demonstrated the similarity of jump height across jump types using a factor analysis in which the first principal component explained 66.4% of the shared variation. The CMJ jump height had the highest correlation with this component, which they called 'explosive power'. Hence, CMJ jump height has good factorial validity, which suggests it could be used as a practical alternative measure of neuromuscular power for field testing where a direct measurement of peak power with a force platform is not available. However, jump height would only serve as a surrogate indicator of neuromuscular power and may respond in a different way to peak power. Confounding factors such as fatigue or training-induced changes in countermovement

depth could potentially de-couple the link with peak power (Cormack, Newton, & McGuigan, 2008; G. Markovic et al., 2011; S. Markovic et al., 2013).

Vertical jumps produce good intra-day and inter-day reliability across a range of kinetic and kinematic parameters, making them suitable for regular athlete monitoring and testing (Cormack, Newton, McGuigan, et al., 2008). Relative peak power and jump height were reported to have inter-day CVs of 3.0% and 5.0%, respectively, which was slightly lower than their corresponding intra-day reliability of 5.2% and 3.6%. In comparison, mean power (absolute and relative) had a higher inter-day CV of 5.5% and 5.7%, respectively, while peak force had a lower CV of 2.2%. Taylor et al. (2010) also found that peak power was a more reliable measure than jump height based on intra-day CV's (3.4% vs. 6.6%, respectively), combined with a smaller worthwhile change (SWC = 3.9% and 4.3%, respectively). Hence, an individual's peak power varied less from jump to jump than it did for jump height. Peak power had a smaller CV than the SWC, which was not the case for jump height.

What becomes clear from these studies is that peak power has marginally better reliability than jump height. Practitioners may consider jump height to be a convenient proxy for neuromuscular power in field-based testing, based on the methods currently available. However, jump height only explains between 72% and 86% of the variance in peak power (Dowling & Vamos, 1993; Harman et al., 1990; Perrine et al., 1978). Jump height is the outcome of the whole jumping movement through its direct relation with impulse, assuming the flight-time definition. In contrast, peak power specifies the maximum output achieved during the jump, which has a more complicated relationship with VGRF, being the product of instantaneous force and velocity (proportional to impulse). Hence, the two measures are of a different nature, so jump height cannot serve well as a proxy of peak power. If a more convenient way can be found to measure peak power, then practitioners may wish to adopt it instead. Practitioners may tolerate a modest error in return for a cheaper, more convenient field-based test. If so, then a suitable target for this new method would be predictive errors below 1.75 $W \cdot kg^{-1}$, the intra-day variance averaged over four studies (Cormack, Newton, McGuigan, et al., 2008; Hori et al., 2007; McLellan et al., 2011; Taylor et al., 2010). This predictive

error will add to the natural variation in an individual's performance from jump to jump. Hence, ideally the model's error should be much lower than this.

2.2.3 Monitoring athlete training status

Jump testing often plays an essential role in athlete monitoring, as part of a battery of tests which typically also include sprint tests, self-administered questionnaires and tests for biochemical markers taken from samples of blood or saliva (Cormack, Newton, & McGuigan, 2008; McLean et al., 2010; Twist & Highton, 2013). If an athlete does not recover adequately from heavy training or a busy competition schedule, their performance may be impaired, or they may become more prone to overuse injuries (Barnett, 2006). Jump performance measures may be more sensitive to fatigue than other tests. For instance, fatigue was still evident in the suppressed jump heights of elite female soccer players up to 69 hours after a match, unlike other measures, including sprint performance, perceived muscle soreness and biochemical markers (Andersson et al., 2008). CMJ performance metrics (flight time and relative peak power) and psychometric tests were found to be more suitable for monitoring professional rugby league players' training status than perceived muscle soreness and tests for testosterone and cortisol levels (McLean et al., 2010). In that same study, there was evidence that relative peak power may be suitable for monitoring 'long term, low-frequency fatigue.' Through a series of CMJ tests before and after an Australian Rules Football match, Cormack, Newton & McGuigan (2008) reported substantial drops in relative peak power and mean power immediately after the match and for a prolonged period afterwards (up to -12.4% and -13.0%, respectively), only recovering after 96 hours. In contrast, jump height dropped more modestly (-3.6%) and had returned to normal after 72 hours (Cormack, Newton, & McGuigan, 2008). This finding suggests that athletes can adapt their jumping action to achieve the same jump height outcome as before, despite their neuromuscular output still being compromised. In summary, peak power is more sensitive to fatigue than jump height, although other measures related to movement strategy could be more sensitive.

2.2.4 Associations with sports performance

CMJ performance measures have a strong relationship with sprint performance, a critical aspect of many different sports (Cronin & Sleivert, 2005). Significant correlations between jump height and sprint times have been reported over various distances (10 m, 30 m, 60 m and 100 m) for elite rugby and football players and young male sprinters, ranging from r = -0.56 to -0.72 (Cronin & Hansen, 2005; Smirniotou et al., 2008; Wisloff, 2004). Stronger correlations were seen in elite male sprinters, from r = -0.77 to -0.93 (Loturco et al., 2015; Young et al., 1995). The findings are not always consistent for peak power as some studies employed different exercises to measure peak power. For instance, absolute and relative peak power in loaded squat jumps were significantly correlated with 40 m sprint times ($-0.76 \leq r \leq -0.52$), but for 10 m sprint times only relative peak power had a significant association (Baker & Nance, 1999a). Relative peak power in hopping was significantly correlated with sprint acceleration (r = 0.80) but not maximal track running velocity (Chelly & Denis, 2001). The authors believed power output was the more influential factor when overcoming the inertia at the start than later in the sprint when running speed was approaching maximum.

While sprinting has a direct correlation with jump performance, there is weaker evidence linking jump performance with broader measures of success in sports. In judo, CMJ jump height and peak power output were able to discriminate clearly between advanced and novice judokas (Detanico et al., 2016). There was a significant difference in relative and absolute peak power between male elite and sub-elite volleyball players, but not in terms of jump height (Sheppard et al., 2008). However, CMJ jump height could not generally differentiate between soccer players from different age groups or between teams from other divisions (Arnason et al., 2004; Castagna & Castellini, 2013; Mujika et al., 2009). There was greater differentiation in motor coordination tests than in jump height between elite and sub-elite female volleyball players in a sport where there is a premium on motor skills (Pion et al., 2015). Finally, American football teams see vertical jump performance as a good indicator of a player's future potential given that it is a prominent factor dictating the draft order in the NFL Combine (Kuzmits & Adams, 2008; McGee & Burkett, 2003).

No one measure of athletic performance can be predictive of success in many diverse sports. Still, it is notable that vertical jump performance, in particular peak power, does appear over and over again in the literature as a means of testing athletes. Their ability to generate neuromuscular power play an important role in all sports to varying degrees, perhaps because jumping is a fundamental skill of human movement.

2.2.5 Relevance of peak power to human movement

The mechanical definition of power is the rate of doing work. This thesis is concerned with the estimation of external mechanical power generated by the body as a whole rather than internal muscular power, which flows from one joint to another in the course of a movement (Knudson, 2009). However, care should be taken when interpreting neuromuscular power as the body's maximal capacity or an intrinsic attribute of the athlete. This notion comes from mechanics, where the power of an engine or motor defines its working capacity. This analogy is a reasonable one for sustained activities such as pedalling on a cycle ergometer or performing repeated jumps (Bosco et al., 1983). In those cases, power production and dissipation occur concurrently, as it does with a mechanical engine (Adamson & Whitney, 1971). However, it is not so apt for an impulsive movement, such as jumping, because power dissipation mainly occurs *after* power production (Adamson & Whitney, 1971). It is more appropriate to think of jump performance being defined by the impulse, which determines the take-off velocity (Knudson, 2009). In this respect, the vertical ground reaction forces (VGRF) characteristics are of prime interest. They represent the net result of coordinated muscular action that 'explains rather than describes the performance' (Winter, 2005). These characteristics of the VGRF curve or their corresponding features of the accelerometer signals may hold the key to predicting peak power using machine learning models.

Several authors have questioned why there should be such emphasis on power when the impulse-momentum relationship determines jump performance (Knudson, 2009; Ruddock & Winter, 2016; van der Kruk et al., 2018; Winter, 2005). However, the relevance of peak power to impulsive actions becomes apparent when considering the principle of late force development proposed by Hochmuth & Marhold (1977). Their

paper showed that over a fixed distance, the best strategy for achieving the highest final velocity was to generate the highest force later rather than earlier in the movement. If force production peaked too soon, then the remaining distance would be covered too quickly without the body maximising its full capability. Their analysis was based on accelerating a point mass with different force-time profiles. It applies to vertical jumping as the point mass represents all the body's mass concentrated at the centre of mass (CM), the final velocity is the take-off velocity, and the fixed distance is the push-off height. The coordinated jumping action can be seen as an optimal control problem subject to constraints where the objective is to maximise the vertical velocity at take-off (Mandic et al., 2016; van Ingen Schenau, 1989; van Soest et al., 1994; Zajac, 1993).

The fixed distance constraint plays a key role in jumping coordination. The body must control the proximo-distal sequence of joint rotations to avoid taking off prematurely before the leg extensors have released all their energy (Bobbert & van Ingen Schenau, 1988; van Ingen Schenau, 1989; van Soest et al., 1994). As part of this, biarticular muscles ensure power is transported efficiently from joint to joint to maximise the external rate of work (i.e. maximise mechanical power) (van Ingen Schenau, 1989). The locomotor system appears to be specifically organised to generate high levels of force late on in the jumping movement (van Soest et al., 1994; van Soest & Bobbert, 1993). By definition, generating the maximum force later in the movement when velocity reaches its highest level requires considerable power.

If jump test performance is to be a useful surrogate measure of general sports performance, then the power-generating locomotor system is of prime interest. The jump height is insensitive to variations in countermovement depth, but peak power output is more dependent on this fixed distance (Mandic et al., 2015). A deeper countermovement would tend to reduce peak power output, but the jump height may be unaffected. Ultimately, it is not the ability to jump high that is of interest to practitioners but how much power is developed in doing so, except for sports such as volleyball where jump height is a critical factor in performance. The assumption being made is that this power-generating capacity has wider applicability, as shown in the correlations with sprint performance and general sports performance above. Thus,

peak power has relevance not just to jumping but to throwing, cutting manoeuvres and many other such sporting movements.

2.2.6 Summary

Jump testing is an important element of athlete monitoring programmes in many professional sports and in academic research. Jump height is a popular measure, yet peak power is more reliable and sensitive to residual fatigue and performance improvements. Given its advantages, if the measurement of peak power could be made more convenient, its benefits could be put in the hands of a broader range of practitioners, far beyond the well-funded professional teams in sports that can afford force platforms. Jumping may not always be a good indicator of overall sports performance, but it can provide insights into sprint acceleration and performance, which plays a critical role in many sports.

2.3 Methods

2.3.1 Gold standard methods

Force platforms have long been recognised as the gold-standard for jump performance measures provided the correct procedures are followed (Owen et al., 2014; Street et al., 2001; Vanrenterghem et al., 2001). The body does work to raise its centre of mass, which is computed from the force platform's measurement of the vertical component of the ground reaction force (VGRF). The peak vertical displacement of the body's CM is regarded as the jump height in this thesis. This definition reflects the work done in raising the body's mass by this height. Some authors use the flight time definition of jump height instead, which ignores the height gained prior to take-off. The external mechanical peak power determined from the VGRF is often normalised to the body mass and called the relative peak power. Henceforth, unless otherwise stated, peak power shall refer to relative peak power. Force platforms are expensive, cumbersome, and entirely unsuited to practical, field-based testing. Hence, a considerable amount of research has been devoted to alternatives.

2.3.2 Power prediction equations

Equations have been developed to predict peak and average power in vertical jumping that were typically based on the body mass and jump height (Table 2-1). Jump height was obtained from the flight time, which can be obtained with a contact mat or an optoelectronic system: $h = \frac{1}{8} gt^2$, where g is the acceleration due to gravity and t is the flight time. The calculation assumes the same degree of leg extension on take-off and landing (hip, knee and ankle), although video evidence shows that participants flex their legs more on landing (Aragón, 2000; Kibele, 1998). The computed jump height is less than the height obtained using the work done method as it does not take into account the rise in the body's CM before take-off. The Sargent jump avoids this problem by measuring the gain in height directly with a jump-and-reach test where the participant makes chalk marks on a wall or displaces vanes of a Vertec device. Buckthorpe et al. (2012), among others, evaluated these different methods with 40 healthy young adults against the criterion of a force plate embedded in the floor of their laboratory. Vertec had a bias of -2.4 ± 6.6 cm (mean ± limits of agreement), which the authors attributed to the skill needed in displacing the vanes at the highest point. The contact mat had the largest bias (-11.7 ± 6.4 cm), which mainly reflected the fundamental difference between the flight-time and work done methods.

Harman (1991) was the first to develop equations to predict peak power from jump height and body mass. His validity measure was based on the Pearson correlation coefficient ($r^2 = 0.77$ for peak power and $r^2 = 0.53$ for average power). He dismissed the original Lewis formula (Fox & Mathews, 1974) for its mistaken assumptions and large underestimations of peak and average powers (by 70.1% and 12.4%, respectively). D. Johnson & Bahamonde (1996) investigated whether other predictors should be included in the equation, finding that the participant's height was also a significant factor, but sex was not. However, later researchers would not have standing height in their models. Sayers et al. (1999) developed equations for average and peak power estimates in the CMJ and SJ based on a large sample size of 108 men and women to produce equations that would be more accurate and more widely applicable. Compared to Harman's SJ peak power equation, the Sayers equation achieved a higher

correlation (r^2 = 0.88 vs. 0.77) and set the benchmark for further research. As Table 2-1 shows, later studies reported r^2 ranging from 0.77 to 0.95, with one outlier of 0.53.

Some of those studies also sought to evaluate previously published equations using their own data. The systematic bias in the peak power estimate for the Harman equation ranged from +4% to -37%; for the Sayers equation from +11% to -27%; for the Canavan equation from +2% to -39%; and for the Lara equation from +23% to -21% (Amonette et al., 2012; Canavan & Vescovi, 2004; Lara et al., 2006; Lara-Sánchez et al., 2011; Quagliarella et al., 2011). The wide variation in errors led some researchers to develop power prediction equations with the express intention of doing so for a specific population (Lara-Sánchez et al., 2011). However, such an approach is not foolproof because the source of the bias may be unknown. Bias may not lie in the participants' sport, age or sex, but it could arise from many other factors. Tessier et al. (2013) concluded that the errors reported were generally too large for such equations to be practical. Knudson et al. (2009) were more critical, pointing out the limitations of such an approach and highlighting the confounding factors of variations in technique and differences in body size (echoing D. Johnson & Bahamonde, 1996). They also pointed out the weak association between power development and impulse, which would be partly addressed by Samozino et al. (2008) (Section 2.3.3). These confounding factors will have to be addressed by any system seeking estimates of peak power by indirect means, as is the intention of this thesis. The machine learning approach will have to be sensitive to different movement strategies, building in other factors beyond the flight time or impulse.

Table 2-1. Power estimation equations from the literature.

Authors	*N*	Sex (Age)	Jump Type	Arm Swing	Equation	r^2	SEE (W)
Harman et al. (1991)	17	M (28.5 ± 6.9)	SJ	Yes	$61.9h + 36.0m - 1822$	77%	Not Reported
Johnson & Bahamonde (1996)	118	M+F (19.6 ± 1.2)	CMJ	Yes	$78.6h + 60.3m - 15.3H - 1245$	91%	462
Sayers et al. (1999)	108	M+F (21.3 ± 3.4)	SJ	Yes	$60.7h + 45.3m - 2055$	88%	372
Sayers et al. (1999)	108	M+F (21.3 ± 3.4)	CMJ	Yes	$19.3h + 48.9m - 2007$	78%	562
Canavan & Vescovi (2004)	20	F (20.1 ± 1.6)	CMJ	No	$65.1h + 25.8m - 1413.1$	92%	121
Lara et al. (2006)	161	M (19.0 ± 2.9)	CMJ	No	$62.5h + 50.3m - 2184.7$	N/A	247
Quagliarella (2011)	117	M (13.6 ± 2.4)	CMJ	No	$61.2h + 52.3m - 1707.7$	89%	415
Lara-Sánchez et al. (2011)	456	M (14.1 ± 0.8)	SJ †	Yes	$61.8h + 37.1m - 1941.6$	N/A	323
Lara-Sánchez et al. (2011)	465	F (14.1 ± 0.9)	SJ †	Yes	$31.0h + 45.0m - 1045.4$	N/A	257
Amonette et al. (2012)	415	M (15.4 ± 2.6)	CMJ	No	$63.6h + 42.7m - 1846.5$	92%	251
Tessier et al. (2013)	80	M (29 ± 7)	CMJ	Yes	Not Reported	95%	231
Tessier et al. (2013)	20	M+F (23 ± 3)	CMJ	No	Not Reported	53%	512
Ache-Dias et al. (2016)	309	M (24.5 ± 5.1)	CMJ	No	$59.5h + 45.8m - 2303.0$	81%	277

h = vertical jump height (cm); m = body mass (kg); H = standing height (cm); SEE = standard error of the estimate; † Abalakov jump

It will be more appropriate to compare the models developed in this thesis with the Standard Error in the Estimate (SEE) reported in the above studies, which is equivalent to RMSE. However, all the equations above predicted *absolute* peak power. In keeping with the sports biomechanics literature reviewed above (Section 2.2), it would be preferable to have an equivalent estimate of the *relative* peak power from these studies to facilitate comparisons later in the thesis. A representative approximation of the relative peak power can be obtained by dividing the SEE by the mean peak power across the cohort. For the studies in Table 2-1, the approximate error in relative peak power ranged from 1.8 $W \cdot kg^{-1}$ (Canavan & Vescovi, 2004) to 7.8 $W \cdot kg^{-1}$ (Quagliarella et al., 2011), with a mean and standard deviation (between studies) of 4.9 ± 1.7 $W \cdot kg^{-1}$. The smallest error of 1.8 $W \cdot kg^{-1}$ came from a sample size of only 20 (Canavan & Vescovi, 2004), which could be heavily influenced by one or two results and may not represent that equation's predictive error in general. This study was considerably smaller than the others (Table 2-1), where the largest involved 456 participants (Lara-Sánchez et al., 2011). Therefore, it would be reasonable to give more weight to the reported errors from studies with larger samples. Rather than weight the studies by N, it would be more appropriate to use $\sqrt{N}$ to reflect how uncertainty in a statistical measure (standard error) scales with sample size. Accordingly, the sample-weighted predictive error for relative peak power is 5.1 $W \cdot kg^{-1}$ across all the results listed in Table 2-1. Restricting the calculation to results for the CMJ without arm swing, the sample-weighted predictive error was slightly lower at 4.6 $W \cdot kg^{-1}$. This figure is a more appropriate benchmark for comparisons with the models developed in this thesis.

2.3.3 Specialised power prediction equations

The above power prediction equations were based on a simple linear statistical relationship and lacked a theoretical rationale. Recognising this, Samozino et al. (2008) derived equations from biomechanical principles for the average force, velocity and power during the propulsion phase of an SJ without arm swing. As with their predecessors, Samazino's equations required the body mass and jump height, but they also included a third term, the push-off distance (the vertical distance travelled by the greater trochanter from the 90° squat position to the take-off position). It enabled the

equations to distinguish between jumpers with the same mass achieving the same height, as their jumps involved different push-off distances. Based on the 11 healthy male participants, the average power equation explained 96% of the variance with a random error of ± 39.8 W. Jiménez-Reyes et al. (2017) applied the equations to the CMJ by providing verbal feedback to the participants to ensure they reached a consistent countermovement depth. The average power's random error was 22.5 W, equivalent to 1.0%, based on a sample of 16 national or international-level sprinters and jumpers. Compared to the results from the simpler equations above (Section 2.3.2), these estimates were more accurate, even after making allowance for peak values being larger than averages, although the samples were much smaller. However, the greater concern lies in controlling the push-off distance as it may depart from the individual's preferred countermovement depth. This may not be an issue at one level since athletes can generally jump from different squat depths despite not having practised them (Bobbert et al., 2008). However, when participants intentionally alter their countermovement/squat depth, peak power output drops when jumps are performed from deeper positions (Kirby et al., 2011; Mandic et al., 2015, 2016). Conversely, jump height is relatively unaffected by such adjustments provided the jump is maximal (Bobbert et al., 2008; Domire & Challis, 2007; Mandic et al., 2015; Salles et al., 2011). In conclusion, Samozino et al. (2008) succeeded in improving the jump test's accuracy, but in so doing the researchers had to switch from measuring peak power to average power. They also required the participants to adopt a controlled squat/countermovement depth that may differ from an individual's preferred level, thereby undermining the test's validity.

2.3.4 Sensor-based methods

The advent of wearable sensor technology prompted several studies into whether sensors could be used to provide accurate estimates of jump performance. Most studies of this sort sought to determine jump height. Although that is not the measure of interest for this thesis, it is worthwhile reviewing the efficacy of these methods as they face the same difficulties and depend on similar techniques that have also been used to estimate peak power output. The validity and reliability of jump height estimates in

the SJ or CMJ vary considerably between studies and the type of method used (Table 2-2). Only in a few cases were random errors less than 10%, and in many cases, they were substantially larger. The systematic errors are not a concern as an appropriate adjustment can be made. Jump height was computed from the sensor's determination of flight time using a proprietary algorithm. Three other papers from Table 2-2 reported an alternative result based on the sensor's estimate of the vertical take-off velocity. Close inspection of the methods described in each paper does not reveal any apparent differences that might account for the superior results in some cases but not others.

The hip-mounted Myotest sensor was responsible for the two best results – random errors of 8.4% (Castagna & Castellini, 2013) and 8.6% (Casartelli et al., 2010) – but it could also produce highly inaccurate results: 22.5% (Monnet et al., 2014) and 71.3% (Choukou et al., 2014). The differences may be due to the precise experimental procedures, such as the tension applied to the Velcro belt holding the Myotest sensor to the body and the participants' individual body composition. Compression strapping, pressing the accelerometer onto the body to hold it firmly in place, can reduce resonant oscillations from soft tissue (Ziegert & Lewis, 1979; Lafortune et al., 1995; Forner-Cordero et al., 2008). Oscillations in the signal were present, as Monnet et al. (2014) reported, which may have been a factor affecting the jump height estimates. The belt would presumably not be adjusted between trials involving different jump types, which may account for the similar errors between SJ and CMJ.

Table 2-2. Summary of studies reporting jump height estimates using accelerometer-based methods.

Study	Participants	*n*	Sex (Age)	Accelerometer	Anatomical Location	Criterion	Jump	Method	Systematic bias ± random error †	Reliability ICC
Casartelli et al. (2010)	Basketball players	44	M (15.3 ± 3.8)	Myotest	Hip	OJ	SJ	V	19.2 ± 42.9%	0.64 $_{(3,1)}$
							CMJ	T	22.0 ± 8.6%	0.98 $_{(3,1)}$
							SJ	T	23.2 ± 8.4%	0.98 $_{(3,1)}$
							CMJ	V	28.9 ± 36.4%	0.75 $_{(2,1)}$
Picerno et al. (2011)	College students	28	M+F (25 ± 2)	Sensorize	LB	MC	CMJ	T	1.7 ± 15.5%	0.89 $_{(3,1)}$
							CMJ	V	35.8 ± 39.5%	0.83 $_{(2,1)}$
Requena et al. (2012)	Professional soccer players	30	M (18.0 ± 2.8)	Keimove	LB	FP	CMJ	T	1.6 ± 14.6%	0.93 $_{(3,1)}$
							CMJ	V	1.2 ± 9.6%	0.94 $_{(3,1)}$
Castagna et al. (2013)	Rugby players	20	M (15.5 ± 0.8)	Myotest	Hip	FP	CMJ	T	-12.4 ± 8.4%	0.88 $_{(2,1)}$
Monnet et al. (2014)	Physical education students	30	n/a (23 ± 2)	Myotest	Hip	FP	CMJ	T	15.7 ± 22.5%	0.67 $_{(3,1)}$
							SJ	T	13.2 ± 23.8%	0.74 $_{(3,1)}$
							CMJ	T*	1.6 ± 21.2%	0.92 $_{(3,1)}$
							SJ	T*	4.0 ± 22.7%	0.89 $_{(3,1)}$
Choukou et al. (2014)	Physical education students	20	M (27 ± 6)	Myotest	Hip	FP	CMJ	T	10.0 ± 71.3%	0.84 $_{(3,1)}$
							SJ	T	15.6 ± 63.6%	0.83 $_{(3,1)}$
Lesinski et al. (2016)	Sub-elite soccer players	19	F (14.6 ± 0.6)	Gyko	LB	FP	CMJ	T	-2.6 ± 12.6%	0.87 $_{(2,1)}$

OJ = OptoJump; MC = Motion capture system; FP = Force platform; LB = Lower Back; V = Take-off velocity method; T = Flight time method; * bespoke method (see text); † 95% CI
For comparison, all results were converted to percentages based on the mean jump height of the criterion where the study did not report results in percentage terms.

Table 2-3. Summary of studies reporting estimates of power using accelerometer-based methods. Peak power reported unless otherwise stated.

Study	Participants	*n*	Sex (Age)	Body Mass	Accelero-meter	Location	Jump	Reported Systematic bias ± random error	For comparison Systematic bias ± random error
Crewther et al. (2011)	Weight trained	12	M (28.8 ± 6.8)	86.8 ± 9.2 kg	Myotest	Barbell	SJ 20kg †	141 ± 896 W	1.6 ± 10.3 W·kg^{-1} *
							SJ 40kg †	-180 ± 593 W	-2.1 ± 6.8 W·kg^{-1} *
							SJ 60kg †	-122 ± 610 W	1.4 ± 7.0 W·kg^{-1} *
							SJ 80kg †	23 ± 400 W	0.3 ± 4.6 W·kg^{-1} *
Choukou et al. (2014)	PE students	20	M (27 ± 6)	74.5 ± 7.2 kg	Myotest	Hip	CMJ	1244 ± 1610 W·kg^{-1} *	16.7 ± 21.6 W·kg^{-1}
							SJ	872 ± 1259 W·kg^{-1} *	11.7 ± 16.9 W·kg^{-1}
Giroux et al. (2014)	Trained/untrained	17	M+F (23.7 ± 3.7)	70.2 ± 11.5 kg	Myotest	Hip	SJ	156 ± 342 W ‡	2.2 ± 4.9 W·kg^{-1} ‡*
Mauch et al. (2014)	Border guards	43	M+F (46.7 ± 7.3)	87.7 ± 12.8 kg	Myotest	Hip	SJ	158 ± 605 W·kg^{-1} *	1.8 ± 6.9 W·kg^{-1}
Hojka et al. (2018)	PE students	33	M+F (21.8 ± 1.7)	69.3 ± 6.5 kg	Myotest	Hip	CMJ	845 ± 1227 W·kg^{-1} *	12.2 ± 17.7 W·kg^{-1}

* Converted from the figures reported, based on the participants' mean body mass.
† Loaded SJ using a barbell with accelerometer attached; PE = physical education.
‡ Average power.
In all studies, the criterion calculation was based on VGRF recorded by a force platform.

There are fewer studies using sensors to estimate peak power in jumping, summarised in Table 2-3, in which the random errors varied considerably. The average predictive error was 10.7 ± 6.3 $W \cdot kg^{-1}$ across all those studies. Based on the sample-weighted calculation above (Section 2.3.2), the weighted predictive error was 11.2 $W \cdot kg^{-1}$, for all results in Table 2-3, and 19.4 $W \cdot kg^{-1}$ for the two studies involving the CMJ. The best result – 4.6 $W \cdot kg^{-1}$ (Giroux et al., 2014) – pertained to the mean power rather than peak power, which is inherently a smaller value. Excluding the loaded squat jumps results, which tend to be more accurate with heavier loading, the best result relevant to this thesis is an error of 6.9 $W \cdot kg^{-1}$ ~ 605 W (Mauch et al., 2014). This error is bigger than all of the peak power prediction equations above (Table 2-1). The SJ peak power estimates tended to be less error prone than the CMJ because the velocity could be specified as zero at the start of the propulsion phase. As with jump height estimates, the prediction errors were substantial to the extent that these methods could not be considered for practical use.

2.3.5 Sensor orientation

The difficulty of estimating peak power (or jump height) based on the take-off velocity is in determining the acceleration and velocity in the true vertical direction. It requires transforming sensor measurements from the device's local frame of reference to the global reference frame, which implies having an accurate estimate of the sensor's orientation. A sensor with only an accelerometer onboard cannot do this except when it is stationary, as only then can it determine the gravitational vector (Veltink et al., 1996). It can find its orientation with respect to a reference position using angular velocity data if the sensor also incorporates a gyroscope. The most straightforward procedure is integrating the noisy gyroscopic data over time, but this quickly leads to drift, curtailing the usable time window significantly (Sabatini, 2011). Instead, various techniques have been developed and refined over the years that use the accelerometer and magnetometer readings (if available) to make corrections continually to the orientation estimate.

Kalman filters correct for drift by estimating the direction of gravity from the accelerometer data, but this only works well in quasi-static conditions (Sabatini, 2011).

Filters that also factor in the direction of the Earth's magnetic field make more accurate estimates of the sensor's orientation, but they are prone to localised distortions in the magnetic field from ferromagnetic materials and electrical sources (Bachmann et al., 2004; de Vries et al., 2009). Extended Kalman Filters (EKFs) use a stochastic algorithm that predicts the future state of the system, adjusting the emphasis placed on the accelerometer and magnetometer data for each time interval (Mazzà et al., 2012; Roetenberg et al., 2007; Sabatini, 2006; Yun & Bachmann, 2006). EKFs are well-suited to human movement studies because their limitations are not generally considered a concern in those situations (Sabatini, 2011). The other leading technique is a quaternion-based, complementary filter, which has achieved similar levels of accuracy at low computational cost (Bergamini et al., 2014; Madgwick et al., 2011; Tian et al., 2013).

Validation studies report that Kalman-type filers have an orientation RMSE generally in the region of 1–5° averaged over the recording period for measurements including trunk sway and hip and knee joint angles in walking, and limb orientations in everyday manual tasks (Bergamini et al., 2014; Cooper et al., 2009; Favre et al., 2008; Godwin et al., 2009; Luinge & Veltink, 2005; Mazzà et al., 2012; Shull et al., 2014). Notably, errors increased for faster movements, such as when stacking crates faster or walking more quickly (Cooper et al., 2009; Luinge & Veltink, 2005). Estimates of attitude (inclination angles) had smaller errors than for heading (direction) (Bergamini et al., 2014). This would be more favourable for the CMJ as it is performed in the sagittal plane with attitude changes. Most of the movements studied were moderate activities performed at relatively slow speeds compared to fast, sporting movements (e.g. the fastest treadmill speed was 2.4 $m \cdot s^{-1}$ or 8.6 $km \cdot h^{-1}$; Shull et al., 2017). The one paper that studied vigorous movements (sweeping and table washing) reported substantial errors in the range of 10–20°, with peak errors amounting to 23% of the task's angular range (Godwin et al., 2009).

All such filters work by anticipating future movements, and so there is always a noticeable lag in response to a change of direction as the filter responds (Figure 2-1). Hence, more dynamic movements with higher inertial accelerations can be expected to incur larger errors. This matters for a system that seeks to quantify performance in

a short, dynamic and discrete movement such as the countermovement jump, which involves a rapid change of direction. Computing the vertical velocity to estimate the jump height or peak power requires the integration of true vertical acceleration. Even with small errors, estimated orientation can lead to substantial error in the jump performance metric. Testament to the precision required comes from the gold standard method where great care must be taken in determining bodyweight and the times of jump initiation and take-off (Owen et al., 2014; Street et al., 2001). Although the best orientation algorithms could achieve an RMSE of ~1° in moderate activities, when applied to jumping, they could not reach the level of accuracy needed to estimate sensor orientation (Picerno, Camomilla, et al., 2011). In conclusion, these Newtonian approaches that compute peak power directly do not appear to be a potentially viable solution as they are overly sensitive to measurement errors.

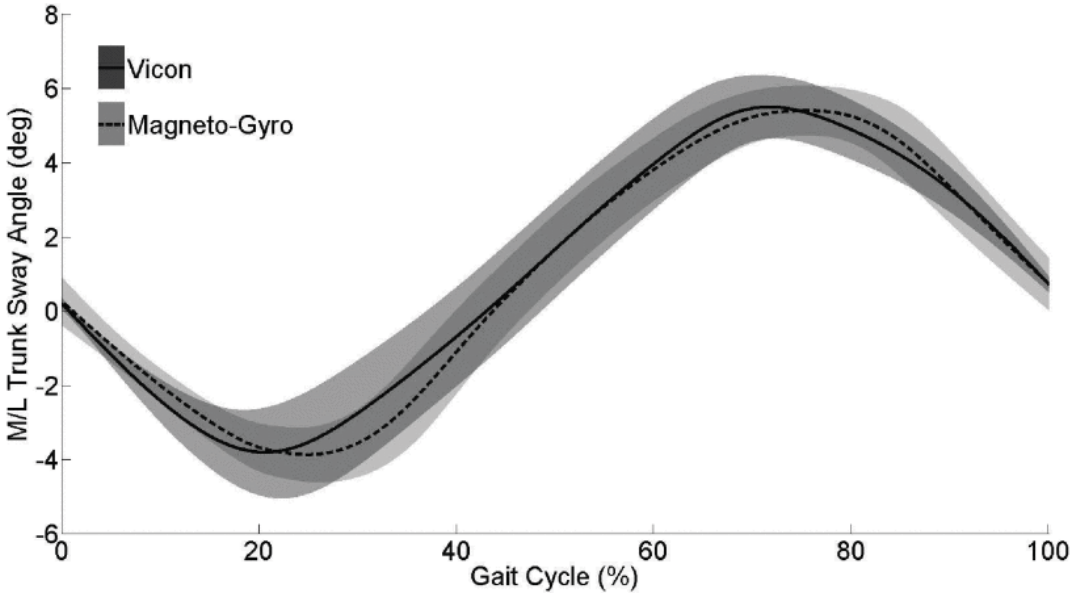


Figure 2-1. Sensor-based orientation lag of magneto-gyro algorithm estimating trunk sway in walking (dashed line) compared to a Vicon motion capture system (solid line). Shading indicates one standard deviation. ©2017 IEEE. Reprinted with permission from Shull et al. (2017).

2.4 Predicting VGRF-related variables

There is a wider body of research into methods to predict the VGRF profile or related discrete measures using wearable sensors as a low-cost, convenient alternative to force platforms. Much of this research relates to gait analysis, where interest focuses on determining peak ground reaction forces or initial loading rates in walking and running related to injury risk (Ancillao et al., 2018). Since peak power is also computed from the VGRF, these studies reveal a wide range of approaches to estimating kinetic variables from sensor data. Moreover, certain aspects of the machine learning methods or their findings have relevance to this thesis as they will help inform the choice of methods. Other techniques using kinematics data from motion capture systems or based on two-segment biomechanical models do not have direct relevance to this thesis and are beyond the scope of this review.

2.4.1 Models predicting discrete VGRF measures

It will be appropriate to consider the studies predicting discrete measures first, as peak power is such a variable (Table 2-4). The early studies in this field established correlations between peak VGRF and the peak resultant acceleration in the CMJ (on landing) with a sensor worn on the upper back ($r = 0.55$, $p < 0.05$) (Tran et al., 2010), or on the shank ($r = 0.90$, $p < 0.01$) (Elvin et al., 2007). Similar correlations were reported for hops, drop landings, and rebound jumps performed by gymnasts ($r = 0.71$–0.83, $p < 0.05$) (Simons & Bradshaw, 2016). The activity tracking studies obtained higher correlations, most likely because they compared values averaged across multiple ground contacts rather than predictions for individual steps (Neugebauer et al., 2012, 2014). Including participant characteristics in the models (e.g. body mass, sex, age), in addition to the sensor data, did not improve their accuracy (W. Johnson et al., 2017; Meyer et al., 2015; Neugebauer et al., 2012, 2014). This finding is consistent with D. Johnson & Bahamonde (1996), who found sex was not a significant covariate in their peak power prediction equation (Section 2.3.2).

Table 2-4. Summary of the key studies predicting GRF-related discrete measures. The text may refer to other studies as this list is not exhaustive.

Study	Sample †	Activity	Measure	Source Data	Sensor Location	Feature Extraction	OT ‡	Model	CV §	*r* / rRMSE *
Elvin et al. (2007)	6 / 18 108	Countermovement Jumps	PVGRF, JH	1A	TB	Peak	N	Correlation	N	0.90, 0.94
Neugebauer et al. (2014)	44 / (6+6)×30s 528×30s	Walking & Running	PVGRF* PBGRF*	3A, SC	HP	Mean Axis Peaks	N	2 × GMM	Y	0.97, 0.66
Simons & Bradshaw (2015)	12 / (10+3+3) 192	Hop, Drop Landing, Rebound Jump	PVGRF	RA	LB, UB	Peak	N	Correlation	N	0.83, 0.73, 0.76
Guo et al. (2017)	9 / 2 18	Walking Tasks	PVGRF	3A	LB, UB, FH	None	N	NARMAX	N	3.8%, 5.0% ¹
Ngoh et al. (2018)	7 / (1+1+1) 21	Walking (10 km·h^{-1})	IVGRF PVGRF	1A	FT	None	N	ANN	N	15.3%, 5.5%
Wouda et al. (2018)	3 / 3×3mins 9×3 mins	Running	KJA, LDRT, PVGRF	3 × 3AGM	XSens	None	Y	2 × ANN	Y	0.85, 0.75, 0.79
Gurchiek et al. (2019)	15 / (1+1) 30	Sprint Start/ Change Direction	VGRF	3AGM	LB	None	Y	Ridge	Y	0.50, 0.66
Derie et al. (2020)	93 / 16 4037	Running	LDRT	2 × 3A	2xTB	Statistical	N	XGB	Y	0.86, 0.88, 0.97 ²
Pogson et al. (2020)	15 / 40 600	Running	IMP, PVGRF, LDRT	RA	UB	PCA	N	MLP	N	0.36, 0.86, 0.79

† Top line: Number of participants / Trials per participant separated in brackets if more than one activity. Trial duration is stated, if relevant. Bottom line: Total sample size.

‡ Signal orientation transformation to global reference frame; § Cross validation; * Correlation or relative RMSE with criterion in order of the measures listed, except where noted.

¹ RMSE for Controlled speed, Freely-chosen speed; ² Subject-independent (no subject features), Subject-independent (with subject features), Subject-dependent.

Discrete Measures: IMP = Impulse; IVGRF = Initial Peak VGRF (rearfoot strike); JH = Jump Height; KJA = Peak Knee Joint Angle (Flexion Extension); LDRT = Peak Loading Rate; MVGRF = Mean VGRF; PVGRF = Peak VGRF; PBGRF = Peak Braking GRF; * indicates the measure was averaged over the trial.

Source Data: RA = Resultant Acceleration; 1A = Uniaxial Acceleration; 3A = Triaxial Accelerations; 3AGM = Triaxial Accelerometer, Gyroscope and Magnetometer (i.e. IMU); XSens = Full-body suit incorporating 17 IMUs (Roetenberg et al., 2013); Underline indicates location that produced best results in study.

Sensor Locations: FH = Forehead; HP = Hip; LB = Lower Back; LPF = Low Pass Filtering; SC = Subject Characteristics; UB = Upper Back (Trunk).

Models: ANN = Artificial Neural Network; MM = Mixed Linear Model; GMM = Generalised Mixed Linear Model; MLP = Multi-Layer Perceptron Neural Network; Ridge = Ridge Linear Regression; NARMAX = Non-linear Auto-Regressive Moving Average model with exogenous inputs; XGB = Gradient-Boosted Tree.

Later studies employed more sophisticated machine learning approaches for walking and running, but the fit was not necessarily better, which may be due to the activities themselves being more dynamic and variable (Derie et al., 2020; Pogson et al., 2020; Wouda et al., 2018). Certain variables may also be harder to predict than others, such as impulse and the peak loading rate compared to the peak VGRF (Pogson et al., 2020). Those kinetic metrics were derived from the predicted VGRF curve that achieved a reasonably good fit, but the plots show large errors (Figure 2-2). The impulse predictions were much less accurate than the other two measures because error accumulated across the time domain. Therefore, it may be better to calculate the measure directly rather than infer it from a predicted curve. This observation has relevance to the aims of the current thesis as peak power is attained immediately following a rapid rise in force production before the propulsion phase.

Given the previous discussion on sensor orientation (Section 2.3.5), it is notable that only two studies chose to reorientate their sensor data to the global reference frame (Gurchiek et al., 2019; Wouda et al., 2018). The study by Wouda et al. (2018) is significant because the authors used the XSens whole-body suit incorporating 17 IMUs, which some authors consider the gold standard for measuring body-segment dynamics with wearable sensors (Shull et al., 2017). They used one neural network to predict joint angles from the XSens orientation data. A second neural network was used to estimate the knee flexion/extension angle, peak VGRF and peak loading rate from the predicted joint angles and the XSens vertical acceleration data. Despite this sophisticated approach using two linked networks, the results (Figure 2-3) were no better than those from a simpler gradient-boosted tree model (Derie et al., 2020) or a perceptron network (Pogson et al., 2020), both of which did not employ a signal orientation transformation.

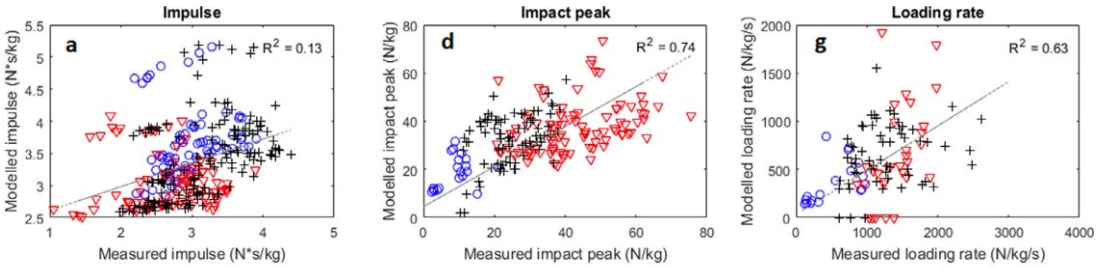


Figure 2-2. Estimation of discrete measures obtained from the predicted VGRF curve for three different activities: steady running (black pluses); running accelerations (blue circles); running decelerations (red triangles). VGRF curves estimated using multi-layered perceptron neural network based on trunk inertial acceleration. Reproduced from Pogson et al. (2020) under open access guidelines.

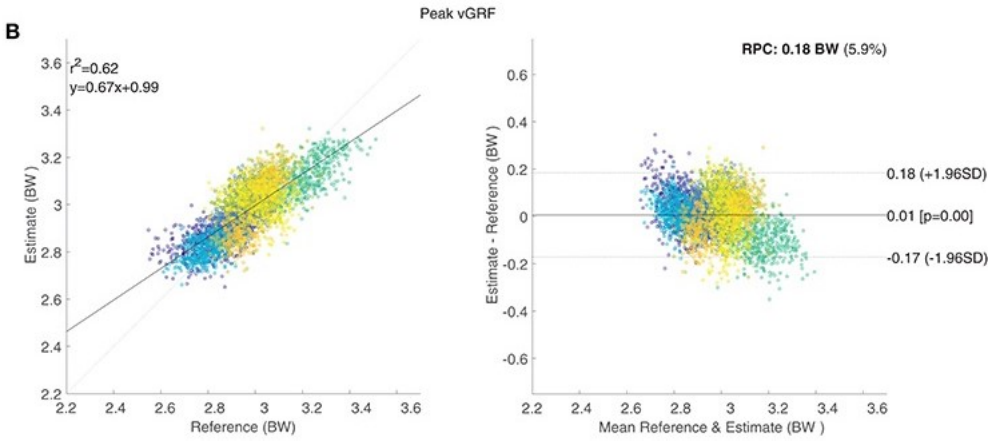


Figure 2-3. Predictions of peak VGRF for each running ground contact-based where the colours identify different participants. Left plot: prediction vs. criterion vGRF. Right plot: residuals (errors) vs. mean of prediction and criterion. The estimates were obtained from two neural networks working in series based on orientation angles and vertical accelerations provided by XSens IMUs. Reproduced from Wouda et al. (2018) under open access guidelines.

2.4.2 Neural networks predicting GRF Profiles

Other studies aimed to estimate the GRF-time profile rather than discrete values, where errors were averaged over the ground contact period. Such reported errors can understate much larger variances seen in certain regions of the curve. A good example of this can be seen in the study by W. Johnson et al. (2019), in which convolutional neural networks (CNNs) predicted VGRF curves for a running sidestepping manoeuvre (Figure 2-4). The overall correlations were high (GRF r = 0.97, 0.96, 0.87,

respectively for the vertical, anteroposterior and mediolateral axes), but there are evidently large variances in the initial loading phase (W. Johnson et al., 2019). The reported overall correlations would have been enhanced by the subsequent longer portion of the signal where the curves were smooth and slowly changing. Moreover, these results were the best achieved with this approach, which was applied to various data subsets. The overall curve fits varied considerably ($r = 0.70–0.96$ for vertical, $r = 0.57–0.95$ for anteroposterior and $r = 0.25–0.87$ for mediolateral) for data sets covering running only, sidestepping only, different running speeds, or running with acceleration or deceleration. W. Johnson et al.'s (2019) correlation method averaged out errors, whereas Pogson et al. (2020) effectively reported a cumulative error.

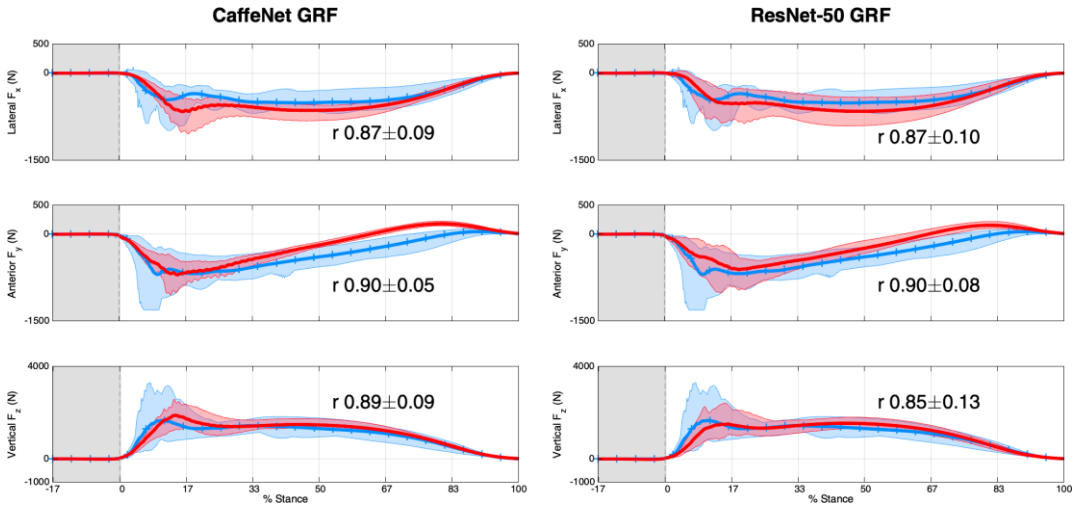


Figure 2-4. VGRF profiles for orthogonal axes predicted by two CNNs using actual triaxial accelerometer data as input (blue = prediction; red = criterion). The networks were trained on synthetic inertial accelerations from motion capture systems. The correlation figures in each of the graphs refer to the range of possible fits obtained with different data subsets. Reproduced from W. Johnson et al. (2019).

It should be noted that the CNNs were trained on synthetic accelerometer data computed from motion capture data drawn from a large database of over 1200 trials. The test data used to produce the plots above (Figure 2-4) were actual accelerometer data provided by another laboratory. The CNNs trained on the resultant acceleration were more accurate than those using acceleration vectors aligned to the principal axis

(anterior direction). Predictions of the Ground Reaction Moment (GRM) were less accurate than GRF with the best correlation of $r = 0.65$, which lends more weight to the notion that measures, such as peak power, derived from the VGRF data may be harder to predict.

Deep neural networks often require substantial volumes of data due to the complexity of their design, the signal-to-noise ratio, and the difficulty of the problem. Across multiple layers and nodes, their numerous parameters amount to a giant regression model with a very large number of degrees of freedom (Hastie et al., 2009). W. Johnson et al. (2019) took advantage of transfer learning to reduce the data set required by using one of two pre-trained networks. Transfer learning is a technique in which pre-trained networks can be re-purposed to new applications (Pan & Yang, 2010). W. Johnson et al. (2019) employed CaffeNet and ResNet-50, two state-of-the-art networks trained on 1.4 million images (1000 classes) from the ImageNet database (Russakovsky et al., 2015). The accelerometer time series were converted into RGB images with a suitable coding scheme, a transformation technique to allow researchers to harness the power of image recognition networks (Z. Wang & Oates, 2015).

An alternative approach would be to use Long Short-Term Memory (LSTM) neural networks, which are specifically designed for time series (Hochreiter & Schmidhuber, 1997). LSTMs were recommended for short activities with a ‘natural order’, such as stepping manoeuvre, based on 4000 experiments with public human-activity data sets (Hammerla et al., 2016). The same study found that CNNs were better for inferring long-term repetitive activities, such as walking or running. LSTMs would appear to be better suited than CNNs to modelling side or discrete movements like a vertical jump. However, pretrained networks are not available because the task itself is quite specialised, limiting any potential gains from transfer learning. LSTMs would typically have to be trained on the data collected in the study. Even the data archive used in W. Johnson et al. (2017, 2018, 2019), which is exceptionally large by biomechanics standards, appears small compared to the ImageNet database. In an earlier study, W. Johnson et al. (2017) trained a partial least squares model (PLS) on the same motion capture data rather than using a CNN. The model produced slightly higher levels of accuracy for the sidestep movement than the CNN for the sidestep

manoeuvre, the only movement considered (GRF r = 0.99, 0.98, 0.97, respectively, as above). The model was considerably simpler, but it still required substantial amounts of data. Individual predictors in the model were mapped to time series points. This arrangement significantly increased the model's degrees of freedom, necessitating a large data set.

It should be remembered that the training data were synthetic accelerometer signals derived from motion capture data, so it was not a true representation. In reality, sensor data is not so extensive as it depends on attachment points. In contrast, motion capture data provides an external view of the activity in the global reference frame. Moreover, if the intended application relates to professional or elite sport, the pool of available high-performing athletes meeting these requirements may be quite limited. That is the case in this thesis as regular athlete monitoring is carried out in professional and elite sport. Therefore, it would be necessary and appropriate to seek an alternative to deep learning networks. A typical machine learning model would have far fewer degrees of freedom as it would depend on a more concise representation.

2.4.3 Feature extraction

It is often necessary to employ data reduction techniques, such as mapping the data onto a lower-dimensional space or computing the various statistical measures, hoping that some of those characteristics may be useful to the model. Finding the right set of features or an effective feature extraction method is vitally important in machine learning (Halilaj et al., 2018). Although activity recognition systems use feature extraction (Cust et al., 2019), only Derie et al. (2020) used statistical measures from the VGRF-related studies listed in Table 2-4. The features extracted from the triaxial acceleration time series in each dimension included the mean, maximum, number of peaks and their timing, as well as more complex features such as coefficients for continuous wavelets, Fourier transforms and an autoregressive model.

These features have similarities with those identified in research conducted in the 1980s and 1990s when scientists sought to identify the VGRF curve characteristics in vertical jumping. They hoped the VGRF curves could become a diagnostic tool providing insights into the temporal and kinetic aspects of jumping so that specific

training programmes could be devised (Oddsson, 1987). Early research was promising as ten force-time parameters were identified explaining 73% of the jump height variance, including the first and second VGRF peak values and their relative and absolute timing (Oddsson, 1987). Dowling & Vamos (1993) undertook a more comprehensive investigation, developing a linear model of jump height with kinetic, kinematic and temporal variables as predictors. Only six of the 19 variables considered were significantly related to jump height: peak force, peak positive and negative power, the ratio of positive to negative impulse, and the timings of peak force and peak positive power. Peak power explained the highest proportion of the jump height variance (86%), which on its own was greater than that for any combination of two or three other variables.

Aragón-Vargas & Gross (1997) took the approach a stage further by devising a comprehensive list of 27 variables (kinetic, kinematic and coordinative) relating to jump execution characteristics. The kinematic variables included joint angles at take-off and peak joint angular accelerations (hip, knee and ankle) and average vertical acceleration. The kinetic variables were external peak and mean power, peak joint moments, peak joint powers, and net joint torques (peak values and at the time of joint reversal). Other variables captured certain aspects of coordination including the relative timing of joint reversals, the time difference between the first and last peak net joint torques and the relative timing of the peak velocity differences between proximal and distal joints for each segment. Crucially, all these variables were discrete, many of which came from motion capture data, not just from the force platform. In certain respects, these variables have more in common with inertial sensor data, which detect movement from its net effect on the body's CM.

The models were limited to three variables to prevent overfitting, so several possible jump height models were presented. The best models included various combinations of external peak and mean power, body mass or the average vertical acceleration, explaining up to 91% of the jump height. On one level, the findings demonstrate the importance of external mechanical power in achieving a movement outcome, whether it be jump height or some other sport-related objective. In another respect, the inclusion of the average vertical acceleration is significant for this thesis as it supports

the idea of using an inertial sensor. The other variables relating to joint powers, moments and the timing of joint reversals were weakly correlated with jump height. The body appears to exploit the degrees of freedom available represented by these variables to achieve the same performance outcome. Such movement variability will be an issue for a model based on inertial accelerations as such additional movements, although intrinsic to human movement, could become a confounding factor in the model. On this point, Dowling & Vamos (1993) concluded that the high degree of variability in the VGRF curves' shape made them impractical as a diagnostic tool. However, the nuances in the VGRF curves may reflect localised elements of the locomotor system, such as co-contraction and other self-organising mechanisms (Davids et al., 2003; Glazier et al., 2006). Hence, they may not have a critical bearing on the performance outcome.

2.4.4 Summary

The studies reviewed above achieved moderate to high levels of accuracy, but none stands out. It is difficult to establish which techniques perform better than others (i.e. sensor attachment location, feature extraction method, model or neural network, etc.) because the interaction between such factors is unknown. There are also differences in sample size, population and experimental design. Nevertheless, this review has revealed that inertial accelerations from body segments do have reasonably high correlations with VGRF measures. Hence, machine learning methods could be applied to body-worn sensor data with some success, as the studies above demonstrated. Transformation of the sensor signal was not necessarily beneficial, nor was the inclusion of participant characteristics in the model. Estimating a discrete measure such as peak power should be done directly rather than via a predicted VGRF curve. Despite the considerable attention they receive, deep neural networks do not always outperform classical machine learning models. They also require large volumes of training data which can be challenging to obtain, particularly in the biomechanics field.

The early research with discrete VGRF measures indicates that discrete measures (kinematic, kinetic or coordinative) might not be the best way to characterise the data because they may discard potentially valuable information (Harrison, 2014). They may

also be inadequate for describing coordinated movement that is inherently variable (Glazier et al., 2006). There are other more efficient data reduction methods that capture the essential characteristic features in the data, including Functional Principal Component Analysis. FPCA yields continuous measures that are statistical in nature, describing the modes of variation across a set of curves.

2.5 Functional Data Analysis

FPCA is a popular technique from functional data analysis, a branch of mathematics that represents time-series data as functions, allowing patterns in the data to be analysed as entities in their own right (Ramsay & Silverman, 2005). It has become increasingly relevant to biomechanics research thanks to new technologies producing greater volumes of time series data (Harrison, 2014; Ullah & Finch, 2013). The data points may be discrete, but the values reflect an underlying biomechanical variable that is continuous. Functional Principal Components (FPCs) describe the time-dependent modes of variation about the cross-sectional mean (e.g. Figure 2-5). The associated FPC scores indicate the degree to which each variation mode is present in each curve. FPCA implicitly assumes a mutual interdependence between neighbouring time points, unlike non-functional PCA, which treats each point as being independent of the others (Section 2.5.5). As such, FPCs describe the time-ordered relationships considered so crucial for LSTMs, but without needing to represent the data as long time series or use such neural networks.

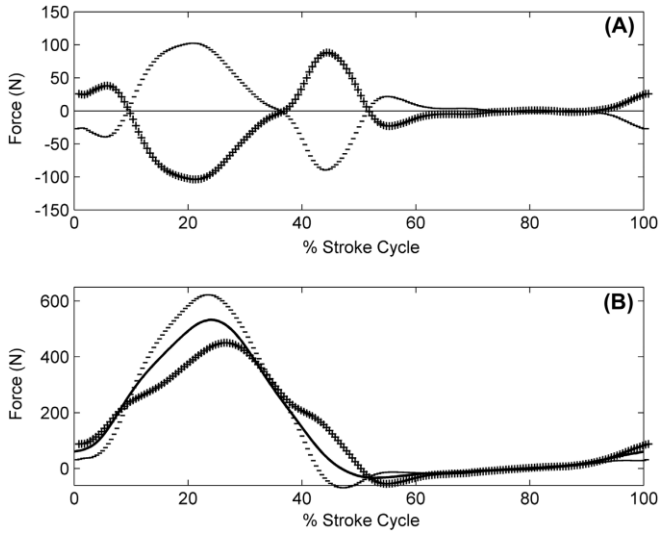


Figure 2-5. Example of a Functional Principal Component: the pin force of a rowing stroke reproduced from Warmenhoven et al. (2017b). (A) FPC function multiplied by twice the SD (pluses) and the negative value (minuses). (B) FPC added to the mean force showing the positive and corresponding negative ranges.

2.5.1 Functional data analysis

Functional data analysis aims to represent a set of discrete observations, y_j, $j = \{1, \ldots, n\}$, according to the model, $y_j = x(t_j) + \varepsilon_j$, where ε_j are the measurement errors and $x(t_j)$ is the quantity of interest. The function x is described by an expansion of K basis functions, $\phi_k(t)$:

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t) \tag{2.1}$$

where c_k are suitably chosen coefficients (Ramsay & Silverman, 2005). B-splines are the most commonly used basis function for non-cyclical data, owing to their flexibility and speed of computation (Ramsay & Silverman, 2005). B-spline functions are piecewise polynomials, defined over a narrow band of the time domain, that overlap with neighbouring b-splines. A b-spline of order n is a polynomial of degree $n-1$, which spans n time intervals. Flexibility can be enhanced by increasing the density of b-splines with more basis functions in the expansion or by using higher order b-splines. For datasets recorded using high sampling frequencies, such as in biomechanics, it is reasonable to use fewer numbers of basis functions than the number of sampling points

provided there remains flexibility to capture the features of interest (Ramsay & Silverman, 2005).

Achieving a good fit for $x(t)$ to the observations, y_j, involves a trade-off between bias and sampling variance. An excessive reduction in bias would have $x(t)$ passing too closely to every observation, making the function follow too closely the noisy deviations in the raw data. On the other hand, reducing sampling variance too far would result in an overly smooth $x(t)$ that does not reflect the true variance in the data. The typical approach (so far as biomechanics is concerned) is to add a regularisation term to the fitting criterion, the residual sum of squares, which penalises high curvatures (second term). The criterion to be minimised, the penalised sum of square errors (PENSSE), is defined as:

$$\mathrm{PENSSE}_{\lambda}\left(x; y\right) = \sum_{j}\left[y_j - x(t_j)\right]^2 + \lambda\int\left[D^m x(\tau)\right]^2 d\tau \tag{2.2}$$

where λ is the roughness penalty and D^m is a differential operator of order *m*. τ is the time as a continuous variable in place of the discrete representation, t_j. Usually, $m = 2$ to obtain the curvature, but higher orders are sometimes required if the differentiated curve is not well-behaved (Ramsay & Silverman, 2005). λ is determined by generalised cross validation (GCV), an efficient optimisation procedure in which a grid search (Section 2.6.3) is often performed to determine largest roughness penalty that does not induce a rise in the GCV criterion (Ramsay & Silverman, 2005). In defining $x(t)$ as a continuous function whilst controlling curvature, FDA assumes that the underlying variable has regularity with no discontinuities, which is appropriate for VGRF and inertial acceleration, the two data sources in this thesis.

The first FPC, $\xi_1(\tau)$, is the eigenfunction that maximises the variance in the corresponding eigenvalues (FPC scores). The FPC scores are defined as the inner product:

$$s_{i1} = \int \xi_1(\tau)x_i(\tau)d\tau \tag{2.3}$$

where s_{i1} is the FPC1 score for the *i*-th curve. Each subsequent FPC (order *m*) is orthogonal to all previous FPCs (and by implication all subsequent ones). Orthogonality requires the inner product to be zero with the further requirement that all FPCs must be normalised:

$$\int \xi_m \xi_{m'} = 0, \int \xi_m \xi_m = 1, \quad m \neq m' \tag{2.4}$$

Hence, the *m*-th order FPC scores are defined as:

$$s_{im} = \int \xi_m(\tau) x_i(\tau) d\tau \tag{2.5}$$

This procedure leads to a progressive reduction in the explained variance associated with each successive FPC. The FPCs, together with the FPC scores, provide an approximation of the original function:

$$x_i(t) \approx \bar{x}(t) + \sum_{m=1}^{p} s_{im} \xi_m(t) \tag{2.6}$$

which can be made more accurate by retaining more FPCs (increasing *p*) from the FPCA procedure. Note that the FPCs define the variance about the cross-sectional mean function, $\bar{x}(t)$. Thus, through this data reduction technique, each curve is defined by a relatively small number of FPC scores, specifying its particular combination of characteristic features. The orthogonality constraint yields FPC scores that are uncorrelated with scores from all other FPCs. Mutual independence is a useful property for predictor variables in machine learning models as it ensures the explained variances in the outcome variable do not overlap. For investigators, interpretation can be made easier by applying a varimax rotation to the FPCs. Varimax rotations rebalance the variance distribution so that each FPC mainly describes a single feature rather than several. The FPCs still retain their orthogonality, but their associated FPC scores become correlated. This transformation is not helpful for statistical models as it introduces multicollinearity to the FPC scores, which becomes a concern when $r > 0.9$ or VIF > 10 (Field, 2013).

2.5.2 Applications to biomechanics

Biomechanical applications of FPCA cover a range of sports and include four investigations of the CMJ (Table 2-5). The articles give prominence to the FPC plots, which present the variance about the mean curve described by a given functional component, offering the non-specialist an immediate and intuitive understanding (Figure 2-5, above). Donoghue et al. (2008) discovered runners who had previously suffered from Achilles tendon-related injuries had movement patterns that were distinctive from those of healthy controls. The authors concluded that FPCA-based continuous kinematic patterns provided more important information than traditional techniques based on discrete variables, such as peak joint angles. Several studies found significant differences in the time-dependent characteristic patterns of the quantities under investigation between different performance levels and other categorical factors (Ryan et al., 2006; Warmenhoven et al., 2018a, 2018b). Bivariate FPCA proved useful as it revealed the interaction between two variables in an intuitive way (Warmenhoven et al., 2017a).

The studies more relevant to this thesis concern regression rather than classification models. In this respect, there are comparatively few regression models in biomechanics, as with the field of machine learning generally. Conveniently, though, those models predicted jump height in the countermovement jump. Peak power has not served as an outcome variable in this context. Richter et al. (2014a) developed a model based on FPC measures computed from the VGRF curves' propulsion phase. It explained 77.9% of the CMJ jump height variance with an RMSE of 2.49 cm. The model fit was much better than when discrete measures were used, similar to those of Dowling and Vamos (1993), where only 20.1% of the jump height variance could be accounted for with a predictive error of 5.03 cm. Moudy et al. (2018) also developed a model from VGRF data using a related technique called Analysis of Characterising Phases, which uses the FPCs as the starting point (Section 2.5.4). The model explained 86% of the jump height with a mean absolute error of 1.39 cm. It was based on the full VGRF curves, from jump initiation to take-off, subject to curve registration (Section 2.5.3). These results are promising, showing the benefit of representing the data in a more conducive form for the model and the outcome variable in question.

Table 2-5. Summary of biomechanics studies using Functional Principal Component Analysis.

Study	Investigation	Participants	Data	Basis	Regist-ration †	FPCs retained	Statistical model
Donoghue et al. (2008)	Chronic Achilles tendon injury	12 injured & 12 controls	Ankle joint angles	n/a *	No	95% ‡	ANOVA
Donà et al. (2009)	Race walking technique	7 national/international	Knee joint angle/ moment	B-splines *	No	95% ‡	None
Ryan et al. (2006)	Countermovement jumps	25 boys/ 24 girls	Knee joint angle	B-splines	Yes	3	DA
Kipp et al. (2012b)	Power cleans	10 collegiate-level weightlifters	Hip, knee, ankle angles / moments	n/a	No	4	ANOVA
Jensen et al. (2013)	Training intervention on CMJ	10 training/ 10 controls (F)	VGRF	n/a	Yes	n/a	ACP
Richter et al. (2014a)	Countermovement jump	121 healthy male athletes	VGRF & RFD	B-splines	No	99% ‡	ACP & LR
Marshall et al. (2015)	Asymmetries in hops and cutting manoeuvres	20 Elite rugby union players (M)	Joint angles, moments & VGRF	n/a	Yes	n/a	ACP
Moudy et al. (2018)	Landmark registration on countermovement VGRF	53 active males	VGRF	B-splines	Yes	7 + 3	ACP & LR
Warmenhoven et al. (2017a)	Sculling technique (bivariate analysis)	2 scullers national/international (F)	Propulsive pin force, oar angle	B-splines	No	n/a	None
Warmenhoven et al. (2018a)	Rowing technique	27 rowers national/international (F)	Propulsive pin force	B-splines	No	n/a	ACP & ANOVA
Warmenhoven et al. (2018b)	Sculling (asymmetries)	40 scullers (M/F)	Propulsive pin force	n/a	No	5	DA & ANOVA

* Fewer basis functions than data points (where reported); † Landmark registration; ‡ FPCs required to explain the specified percentage of explained variance.
ANOVA = Analysis of Variance; DA = Discriminant Analysis; ACP = Analysis of Characterising Phases; LR = Linear Regression.

These studies show FPCA can be a useful feature extraction technique for discriminating between classes or predicting performance outcomes. FPC scores may be the type of predictors needed for such a predictive model that would take the place of the discrete measures employed in the past. Therefore, a model based on FPCA is worth investigating in this thesis as a feature extraction method for accelerometer data. It has not been applied to wearable sensor data previously in the biomechanics field.

2.5.3 Curve alignment

A requirement for FPCA and FDA, in general, is that the curves should be of equal length in the time domain. In practice, that usually means the times series having the same number of data points across all trials, but that is not a strict requirement as data from different sources may be sampled at different rates. When the data comes from the same source, as is usually the case, the time series from all trials need to be standardised to the same number of points. The simplest method is to use linear length normalisation (LLN), a simple interpolation method that resamples the time domain (Chau et al., 2005). However, such linear transformations can shift the temporal positions of features relative to those same features in other curves, increasing phase variance rather than reducing it (Page & Epifanio, 2007). Consequently, where features do not align, the cross-sectional standard deviations may be inflated over certain periods (Chau et al., 2005). The corresponding FPCs may not be sharply defined as their variance may include an element of phase variance as well as amplitude variance.

This shortcoming can be addressed in FDA with curve registration which aligns common features (i.e. landmarks) using a suitable nonlinear function, $h(t)$, that transforms the time domain: $f(t) \rightarrow f(h(t))$ (Ramsay & Li, 1998). Registration effectively separates the variation between curves into amplitude and phase variance so they can be analysed independently. The landmarks should be clear and unambiguous across all curves, such as maxima, minima or crossings of fixed thresholds, often the horizontal axis (Ramsay & Silverman, 2005). However, whilst registration may ensure the corresponding landmarks line up across the curve set, phase variance can still be in evidence away from the landmarks (Kneip & Ramsay, 2008). Continuous registration can mitigate this when implemented after landmark

registration, but there are no applications in the biomechanics literature (Ramsay & Li, 1998). Moudy et al. (2018) conducted the only biomechanics study into landmark registration, applying it to the VGRF curves from countermovement jumps. The authors reported that using a single landmark (VGRF maximum) was better than more landmarks or none, based on the total explained variance in jump height. From this limited evidence, landmark registration appears capable of marginally improving a linear model's fit, but the number of landmarks should be kept to a minimum.

Some researchers have questioned whether registration should be applied mainly in situations where the interest lies in the time domain (Preatoni et al., 2013). Preserving the time domain is essential for calculating peak power from the VGRF curve, so it may be appropriate to do the same with sensor data. On the other hand, it may be more important to achieve a high degree of alignment with registration if the temporal definition of sensor-based FPCs is less of a concern. There would be a direct association with impulse as there is with VGRF data. Some investigators have used registration to counteract the effects of LLN (e.g. Floría et al., 2016; Warmenhoven et al., 2017b; Moudy et al., 2018). Others have used padding to extend the shorter curves, so all curves were of the same length (Donoghue et al., 2008; Ryan et al., 2006). Vertical jump data (VGRF or sensor data) can be padded at the start without distorting the curves aligned at take-off. The propulsion phase closest to take-off has the least temporal variability, followed by the braking phase immediately preceding it (J. McMahon et al., 2017; Sole, 2015), ensuring a high degree of alignment without intervention.

2.5.4 Analysis of Characterising Phases

An FPC can often describe several features across the time domain, indicating some correlation between them. Interpretation can usually be enhanced with varimax rotations, but multicollinearity may be an issue for statistical models (Section 2.5.1, above). A non-rotation approach is Interpretable FPCA, which seeks to zero FPCs in the regions of the time domain where their respective variances are not significant (Lin et al., 2016). It improves interpretability at the expense of a slight reduction in explained variance, but it has few applications in the literature. Another approach is to

use the Analysis of Characterising Phases (ACP), developed by Richter et al. (2014a), which has become popular in biomechanics (Table 2-5, above).

ACP can improve interpretability based on similarity scores instead of FPC scores, computed from key phases of varimax-rotated FPCs. The key phases are where the FPC is above a certain percentage of its peak value, usually specified by the investigator (e.g. > 90%; Moudy et al., 2018). The similarity scores (ACP scores) provide a measure of how closely aligned an individual curve is with a reference curve, which may be the cross-sectional mean curve (Richter et al., 2014b) or the curve associated with the best performance (Richter et al., 2014a). Taking the individual curve's mean over the phase is a simpler alternative (Richter, personal communication). ACP has been used for classification models (Warmenhoven et al., 2017b) or comparative analysis using *t*-tests (R. Jensen et al., 2013; Marshall et al., 2015), but the prime interest in this thesis is with regression models.

A linear model based on the ACP scores accounted for 98.8% of the jump height variance, compared to 77.9% achieved by the FPC model (Richter et al., 2014b). However, when ACP was applied in a separate study with 53 participants based on VGRF data from the CMJ, the model explained at best 86% of the jump height variance (Moudy et al., 2018). Richter et al. (2014a) only used the jump's propulsion phase, whereas Moudy et al. (2018) included the whole curve from jump initiation to take-off, which may account for the disparity in model fit. Both studies defined jump height based on the take-off velocity, which is directly proportional to the area under the VGRF curve, but it does not take into account the rise in the body's CM before take-off. The regression models amounted to a linear combination of predictors proportional to impulse, which is reflected in the definitions of FPC and ACP scores. Had the jump height been defined according to the work done definition, there would have been no direct correspondence with impulse.

One of the advantages of ACP is its ability to identify variances in the time and time-amplitude domains, thus providing temporal characteristics. The ACP scores included in the models above were based primarily on the amplitude-time domain, reflecting the time domain's importance in an impulse-dependent measure (Moudy et al., 2018; Richter et al., 2014b). However, these ACP scores may have figured prominently

because of how the VGRF curves had been aligned. Much of the temporal variance arose due to differences in jump execution time that is illustrated well in Figure 2-6, which is taken from another ACP study (R. Jensen et al., 2013), although Richter et al. (2014a) had a similar plot. Had the curves been aligned at take-off, then the temporal variances would have been much smaller, and consequently, the time-based ACP scores may not have been included in the model. This example highlights the importance of curve alignment (Section 2.5.3, above).

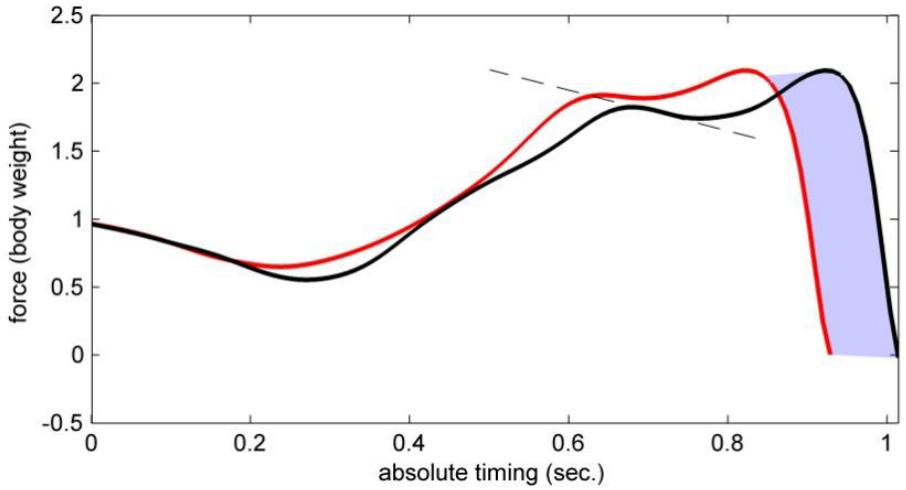


Figure 2-6. Example of VGRF curve alignment at jump initiation, re-produced from R. Jensen et al. (2013). The jump takes longer to execute following a training intervention (black line) than before (red line). The dashed line indicates the start of the propulsion phase.

In conclusion, ACP has outperformed FPCA where a direct comparison was made, and there is evidence that it more efficiently captures the features, requiring fewer ACPs than FPCs (Richter et al., 2013, 2014b). ACP achieved exceptional results in circumstances where the ACP scores corresponded closely to the outcome variable, being equivalent to impulse for a jump height model. However, it is unclear whether the good fit was an artefact of the data or a strength of the method itself. A similar situation would not likely arise if sensor data was used as the input, while peak power was obtained separately from the VGRF data. In conclusion, ACP does not appear to be a suitable choice for sensor data in this thesis.

2.5.5 PCA of waveforms

The characteristics defined by FPCA are intuitive because they describe the shapes of the variation modes extending over the whole curve. An individual curve can be represented in practical terms by a handful of numbers, the FPC scores, that quantify the degree to which each feature is present. In contrast, the (non-functional) PCA of waveforms defines a curve as a full time series in which the PC scores vary as a function of time (i.e. a separate score for each data point) (Deluzio & Astephen, 2007). The modes of variation can still be plotted as a series of points, but no single mathematical entity represents each mode. Typically, a neural network would be required to handle so many inputs, which inevitably entails large volumes of training data that may not always be feasible (Section 2.4.2). However, the concise FPC representation allows classical ML models to be employed if so desired.

FPCA and PCA are different ways of characterising the same data, so they can be expected to produce equivalent results (Warmenhoven et al., 2021). The advantage of FPCA over PCA for machine learning lies in its efficient data representation, simplifying the model and avoiding complexity that may be unnecessary. Nuances in the data are not necessarily lost as they are abstracted into higher-order FPCs, but their relationships with other more prominent features would be lost. However, how the data is represented or how a question is posed can be crucial. In conclusion, FPCA is a useful tool in many applications for extracting essential characteristics from the data. Hence, it would be appropriate to adopt it for this thesis as the preferred method for feature extraction. It has the twin advantages of keeping models relatively simple and the sample size requirements moderate.

2.6 Machine learning framework

Having a model and a means of extracting features from the data are essential elements, although those two elements are not sufficient on their own. Attention should be paid to selecting the appropriate model (including optimisation). In addition, a robust estimate of the model's predictive error is needed for when it is applied to new data. Cross validation (CV) can provide the framework in which model selection and

appraisal can be performed with rigour. However, it is not a procedure that has always been followed in the biomechanics literature, as Table 2-4 indicated.

2.6.1 Cross Validation

The acceptance of a proposed model (or neural network) depends on its ability to make accurate predictions when presented with previously unseen data (Stone, 1974). Cross validation is an empirical and versatile tool for estimating a model's predictive error when applied to unseen data (Allen, 1974; Geisser, 1975; Stone, 1974). CV involves repeatedly splitting a data set in two, training the model on one portion of the data and then testing it on the remainder. Through repetition, a generalised (average) estimate of the model's predictive error can be obtained. Its principal assumption is that test data should be drawn from the same distribution as the training data. It can be applied to a wide variety of models of many kinds as it makes minimal assumptions. As such, it places emphasis on the predictive error of the model rather than on the accuracy of parameter estimates that is a hallmark of traditional statistical inference (Geisser, 1975). As an estimator of a model's predictive performance, CV is one of the most commonly used criteria for model selection (Y. Yang, 2007).

Early CV implementations were based on the Leave-One-Out (LOO) design in which the model is trained on all but one data point and then tested on that single case. The procedure is repeated N times. It has the apparent advantage of yielding almost unbiased error estimates, albeit with substantial variance, but this comes at a high computational cost (Efron, 1983). When evaluated against other estimators based on a logistic regression model, it was outperformed by bootstrap methods which produced errors estimates that had much lower variance with only moderate bias (Efron, 1983, 1986).

Bootstrapping involves randomly sampling from the whole data set with replacement, such that the same case could be selected more than once. The model error for the resampled data is averaged over a larger number of iterations, typically > $2N$ (Shao, 1993). The best estimator was the '0.632 bootstrap', which was a compromise between cross validation and maximum likelihood, albeit with an admittedly weak theoretical justification (Efron, 1986). An improved method (0.632+ bootstrap) outperformed

various CV methods (LOO and K-Fold) in a more comprehensive evaluation involving 24 simulation experiments, although all were classification models (Efron & Tibshirani, 1997). This new bootstrap method was more computationally efficient than cross validation, although this is less of a concern today with the computational resources available on the desktop. However, when applied to real-world data sets for some problems, bootstrap estimates could be extremely biased (Kohavi, 1995). In contrast, the 10-fold and 20-fold CV designs produced more accurate and reliable estimates for the six classification models considered. The preference today for 10-fold CV can be traced back to Kohavi's 1995 influential paper. However, no substantive difference was reported in a later paper between 5-fold and 10-fold CV in simulations of three classifiers, although 10-fold was marginally the better one in terms of its combination of bias and variance (Braga-Neto & Dougherty, 2004).

Model selection has different concerns to the model assessment above as the best estimator is the one that selects the optimal model with the highest probability. An estimator is said to be asymptotically consistent if the probability of selecting the optimal model tends to one in the limit as $N \to \infty$ (Shao, 1993). However, LOOCV is not asymptotically consistent, so there is a risk of choosing a suboptimal model that diminishes with increasing sample size (Shao, 1993). Shao (1993) showed that this shortcoming could be rectified by increasing the size of the validation set to at least half the data set. In 1000 simulation runs of regression models with up to five predictors, LOOCV was among the worst selectors, alongside other estimators of the same type: AIC (Akaike Information Criterion), Mallow's C_P and GCV (Generalised Cross Validation) (Shao, 1997). LOOCV tended to be too conservative, selecting an unnecessarily complex model, or it could not distinguish between under-specified models (Shao, 1993). The best selector was the 'delete-d' CV with a validation set somewhat larger than the training set (25 cases to 15). Such a data split is at the opposite end of the spectrum from LOOCV, which has the smallest possible validation set. The discovery ran counter to conventional thinking at the time. Shao concluded that it was wiser to use a relatively small training set because the prediction error hardly changes for over-specified models. Conversely, if the validation set is too small, it is harder to distinguish between candidate models. He recommended a two-stage process

in which model selection is based on CV with a small training set, followed by the selected model being re-trained on a bigger data set to estimate its predictive error (Shao, 1993).

Shao (1993) used a similar experimental simulation to show that CV designs with many data splits were much more likely to select the optimal model than LOOCV. The Monte Carlo technique (MCCV), which was first proposed by Picard & Cook (1984), was developed further with mathematical proofs and analysis by Zhang (1993), who called it multi-fold cross validation. MCCV involved resampling the data set without replacement, in contrast to the bootstrap that did employ replacement. Unlike the K-fold designs, each resample ('replicate') is independent of the others. In all cases, MCCV had a success rate of > 0.9, whereas LOOCV achieved 0.5 for simple models, which improved with more predictors (Shao, 1993). The results were based on linear models, but the theorems apply to more complicated models, such as nonlinear regression and generalised linear models. MCCV has not been used widely in the literature, where it is also called Leave-Many-Out (K. Baumann, 2003). It has been applied in chemical engineering, where it was more likely than LOOCV to select a model with the correct number of predictors (Xu et al., 2004; Xu & Liang, 2001). It has been noted that there was little further improvement in error estimation from increasing the number of iterations from 20 to 50 to 1000 (Molinaro et al., 2005). An MCCV design at the lower end of this scale is comparable to the repeat K-fold designs that have become popular in recent years (Burman, 1989; P. Zhang, 1993). Such repeat K-fold designs have outperformed ordinary 10-fold CV (Braga-Neto & Dougherty, 2004); and the 0.632+ bootstrap (Kim, 2009). Repeated K-fold CV reduced model internal sensitivity (an estimator's inherent randomness) to an insignificant proportion of the variance (Rodríguez et al., 2013). Therefore, it is recommended as an efficient and reliable alternative for error estimation, even with 10 or 20 repetitions (Y. Zhang & Yang, 2015). However, as the number of repetitions increases, repeat K-fold becomes indistinguishable from MCCV since the distinct validation sets within each repetition increasingly overlap those from other repetitions.

The rise of repeat K-fold designs is indicative of the increasing computing power that is available today. Resources may also be devoted to fitting models with increasing

complexity, such as in DNA microarrays typically with several thousand predictors but only a hundred or so samples (Ambroise & McLachlan, 2002; Braga-Neto & Dougherty, 2004; Simon et al., 2003). Neural networks can be orders of magnitude more complex and require substantial data sets to achieve good results (Section 2.4.2). In those circumstances, it is more common to see holdout validation employed instead in which a portion of the data, typically 10-20% is set aside to test the model independently. Training the network may still involve a limited cross validation design. However, such is the associated computational demands that it may not be feasible to use repeat K-fold, MCCV or bootstrapping. Thus, the CV methods described above tend to apply to machine learning models rather than neural networks. In the latter case, where the computational costs are substantial, a holdout set can be used to test the network independently.

2.6.2 Nested cross validation

Cross validation makes efficient use of the data by repeatedly splitting it into training and validation sets. However, such a procedure introduces selection bias because the overall evaluation is not made independently as the same data is used for training and validation. This influences the choice of features or parameters values, which in turn, biases the selection procedure (Stone, 1974). One approach was to adjust the error estimate, but using a statistical formula (e.g. shrinkage) was considered unsatisfactory (Mosteller & Tukey, 1968). Stone (1974) proposed his double-cross validation procedure in his seminal work, but it had a considerable cost that delayed its adoption for nearly thirty years (Ambroise & McLachlan, 2002; Simon et al., 2003; Varma & Simon, 2006). Today, it is more often known as nested cross validation (NCV), the term that will be used in this thesis.

Nesting, as the name suggests, involves performing cross validation inside two nested programming loops. Model training, feature selection and parameter tuning are performed inside the ‘inner’ CV loop, which sub-divides the training set from the ‘outer’ loop. The selected model that emerges from the inner loop is then re-trained and evaluated on the full training and validation sets at the outer level. The procedure then repeats for each outer iteration. This structure ensures model assessment is kept

separate from all aspects of model selection, but it results in a range of different models, one for each outer fold. However, a final aggregate model can be obtained using the bagging technique that averages the parameter values, producing a more generalised model (Breiman, 1996).

Several studies have demonstrated that without nested cross validation, bias can sometimes be substantial. Varma & Simon (2006) showed non-nested CV error estimates could deviate from the true mean error of 50% by 8–12 percentage points on average in binary classification models. In one fifth of cases, the predictive error exceeded 20 points (Figure 2-7). Other investigators reported K-fold CV procedures could produce overly optimistic estimates that were little better than the training (re-substitution) error (Ambroise & McLachlan, 2002; Simon et al., 2003). Cawley & Talbot (2010) presented a series of examples demonstrating selection bias, which could be severe in some problems, particularly for small sample sizes. The authors described selection bias as a form of overfitting, equivalent to overfitting an over-specified model, but noted that in 2010 there were few examples of NCV in the literature. However, a recent survey of 55 ML papers in autism research revealed an inverse relationship between sample size and classification accuracy (Figure 2-8), suggesting selection bias led to overly optimistic estimates (Vabalas et al., 2019). In the same paper, Vabalas et al. presented a comprehensive analysis of NCV compared to K-fold CV based on high-dimensional classification models (SVM and logistic regression). They showed that selection bias was large, although it diminished with sample size (Figure 2-9). They concluded that although NCV was computationally demanding, it was necessary to prevent overfitting. However, in situations where the cost of NCV would be prohibitive, a single test set could be set aside to provide an independent test so overfitting may be diagnosed (Feurer & Hutter, 2019). It is essential that the holdout set is sufficiently large so the test error has meaningful confidence intervals (Simon et al., 2003). Despite the advances that have been made in cross validation methods, as discussed above, the holdout test still retains its attractions because it is demonstrably a truly independent test of the model.

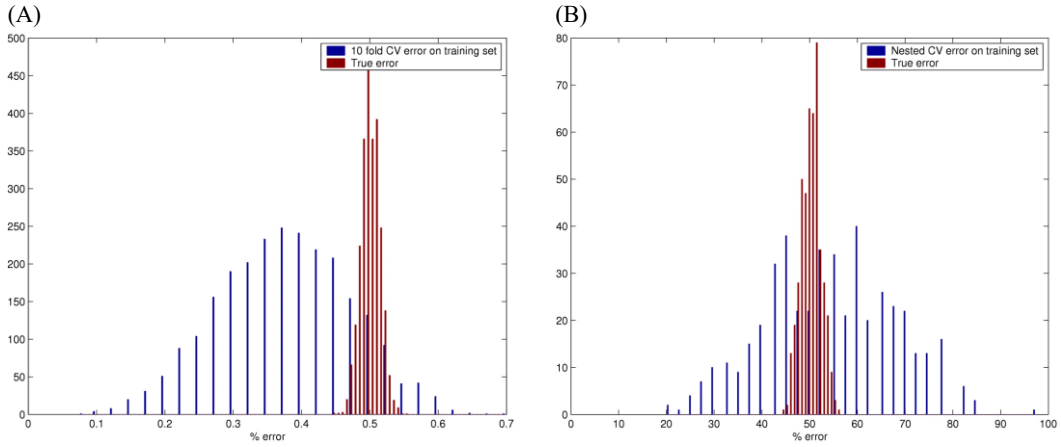


Figure 2-7. Distribution of classifier accuracy for a shrunken centroids model in a cross validation (blue lines) compared to the true error (red lines), based on (A) 10-fold CV and (B) Nested CV. Reproduced with permission from Varma and Simon (2006).

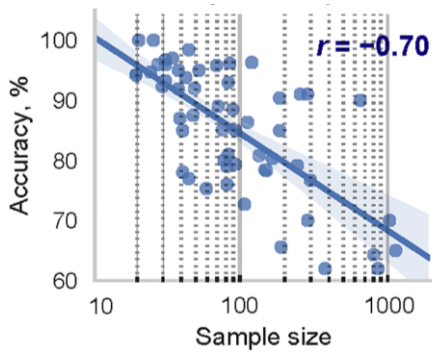


Figure 2-8. The inverse relationship between the reported classification accuracy and sample size from a survey of 55 machine learning papers in autism research. Reproduced from Vabalas et al. (2019) under open-access guidelines.

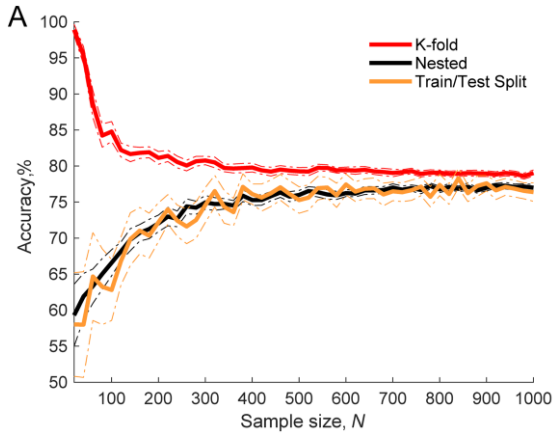


Figure 2-9. Comparison of cross validation schemes for increasing sample size: K-fold CV (red), Nested CV (black), Holdout (orange). K-fold CV substantially over-estimates accuracy for small samples, but Nested CV shows no apparent bias. Reproduced from Vabalas et al. (2019) under open-access guidelines.

2.6.3 Optimisation

Model selection is an optimisation problem in which the objective function is the loss (CV error) that needs to be minimised. In model selection, there is no distinction between hyperparameters governing the algorithm's operation and other parameters having a role outside of the model, such as in data preprocessing (Engels & Theusinger, 1998). The most popular method is to perform a grid search. The algorithm iterates through a range of possible parameter values, evaluating the model at each step (usually CV error) to find the optimal parameter values (e.g. LeCun et al., 1998). Its popularity has been attributed to its simplicity and reliability in one and two dimensions and because it typically finds a better solution than a manual, expert-led search (Bergstra & Bengio, 2012). A random search, on the other hand, has the same practical advantages, yet it is considerably more efficient than a grid search when working in higher dimensions (Bergstra & Bengio, 2012).

A more sophisticated approach uses Bayesian optimisation, which undertakes a directed search based on a Gaussian Process (GP) surrogate model, balancing exploration of the parameter space with over-exploitation of promising regions (Bull, 2011). It has outperformed other global optimisation algorithms in benchmark tests (Snoek et al., 2012). Bayesian optimisation is well suited to machine learning algorithms because they are expensive to evaluate, either by CV methods or by having to train a neural network according to new parameters. It was first proposed by Mockus (1977), but as with NCV, it only became more popular when sufficient computer power became available (D. Jones et al., 1998). The algorithm works best with categorical and few, if any, numerical parameters (Eggensperger et al., 2013). Its performance degrades in higher dimensional spaces, typically beyond 15–20 dimensions, which can be mitigated by using tree-based rather than GP surrogate models or using dropout techniques that exploit the inherently low dimensionality of the parameter space (Eggensperger et al., 2013; Li et al., 2017; Z. Wang et al., 2013). Bayesian optimisers also perform poorly with noisy objective functions, such as CV estimators, but various solutions have been proposed. It remains an active research area that will need to be addressed in this thesis (Gramacy & Lee, 2011; Letham et al., 2019; Picheny, Ginsbourger, et al., 2013; Picheny, Wagner, et al., 2013).

2.6.4 Data augmentation

Machine learning models are limited in their scope and complexity by the availability of training data. A more extensive data set allows more variables to be included in the model, thereby refining model prediction and guarding against overfitting (Bishop, 2006; Hastie et al., 2009). In machine learning, small data sets are those where there are not sufficient samples to support the likely number of predictors required in the model, such as in genetics, where there may be more genes than there are samples (Simon et al., 2003). Cross validation procedures are generally not suitable for small data sets because after splitting the data, the training set becomes too small to cover the variety of possible test data it may encounter. If the validation set is too small, a representative assessment cannot be made of how well the model would perform on unseen data (Shao, 1993). However, it can be challenging to obtain the sample size needed to achieve the accuracy desired in all fields of machine learning, if not more so, when relying on volunteers in the human sciences. Participants have to meet the requirements for the study and be willing to take part. Data collection itself can often be time-consuming, more so for the researcher than the participant, given the setup and subsequent data processing involved. The study itself must be first approved by the relevant authority concerned with the ethics of the investigation. One way to address this problem is to invite participants to perform multiple trials to increase the data set's size, as shown in the studies into predicting VGRF-related measures (Table 2-4). However, the number of trials that can be performed will be limited by practical, biological and ethical considerations. Instead, synthetic data can be generated to augment the training data based on manipulating the existing data set or using an alternative data source. This section highlights well-established augmentation techniques that can be applied in this thesis rather than offering a comprehensive review.

Many data augmentation techniques based on the original data set were developed for classification problems with unbalanced data sets, such as in medicine, where identifying rare conditions was the aim. Training classifiers to identify the minority cases is problematic because there are so few of them, such as in cancer diagnosis (Chawla et al., 2002). The aim of augmentation for this purpose is to rebalance the

data set by generating artificial cases of the minority condition. The SMOTE algorithm (Synthetic Minority Oversampling TEchnique) is a well-established method developed to address this problem using linear interpolation between two real cases in feature space to generate a new synthetic case (Chawla et al., 2002). This idea was developed further with the ADASYN (Adaptive Synthetic) sampling algorithm, which used a weighted distribution to select cases for augmentation. It was extended to regression problems where the aim was to estimate the probability of rare events, such as volcanic eruptions (Torgo et al., 2015; Torgo & Ribeiro, 2007). Augmentation may also be applied directly to the raw data with an appropriate transformation. A key example is from the image recognition field, where a network can be trained on rotated or inverted images in addition to the originals. According to a systematic review, this approach has improved classification accuracy by several percentage points in many cases (Shorten & Khoshgoftaar, 2019). The same principle could be applied to three-dimensional sensor data where small rotations would be equivalent to slight variations in the angle the sensor was attached to the body.

Utility-based regression defines a relevance function for continuous outcome variables that corresponded to the cost-benefit matrices used in classification models (Torgo & Ribeiro, 2007). It can be used to target the synthetic case generation to ranges of the outcome variable where it is important to be accurate (the benefit, given the application) and where the model performs poorly (the cost) (Torgo et al., 2015). The model may be trained on data from participants across the performance range, but all things being equal, there will be fewer examples of jumps from the top end of the scale. Consequently, the model may be less accurate when tested on high performers.

An alternative approach to augmentation is to draw on another data source that can be transformed into the form required for the model. Biomechanics research has a couple of instances of this where kinematic data has been transformed into inertial accelerations as if they had been generated by an accelerometer attached to the body. Mundt et al. (2020) developed this approach to estimate lower limb joint angles and moments for 93 participants walking on a treadmill from accelerometer data from five body-worn IMUs using a neural network trained on real and augmented data computed from a motion capture system. The simulated acceleration data correlated with the

measured inertial accelerations, ranging from $r = 0.95 \pm 0.08$ for the pelvis sensor to $r = 0.88 \pm 0.12$ for the right thigh. Without augmentation, the neural network estimated the joint angle and moments with correlations of 0.85 and 0.95, respectively. When the simulated data were included, the kinematic correlation rose slightly to 0.89, but there was no change in kinetics correlation. The RMSEs dropped slightly in both cases, from 4.8° to 4.3° for joint angles and from 13.0% to 11.6% for moments. Despite a large data set (for the kinematics, there were 3098 actual samples and 46437 simulated), the improvement in prediction accuracy was only marginal. W. Johnson et al. (2019) also used kinematic data from motion capture to simulate sensor-based inertial accelerations for predicting GRF and GRM. However, the authors did not report the benefit of augmentation or break down the results in this way.

More recently, augmentation methods have been developed using generative models including Variational Autoencoders (VAEs) (Kingma & Welling, 2014) and Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). These models are deep neural networks specifically designed to learn nonlinear representations of the data. VAEs randomly draw from the encoded features' distributions to generate synthetic variations on the original data. In contrast, a GAN's generator network has no direct access to the training data, but it learns to produce increasingly realistic synthetic data through a contest with the discriminator. The discriminator trains on real and synthetic data, learning to classify them correctly as real or fake. A minimax loss function incentivises the generator to increase the misclassification rate, while the discriminator tries to lower it. Hence, the generator learns to produce data that increasingly resembles real data, while the discriminator improves it ability to spot the difference. These generative models have achieved considerable success, such as producing fake images that a human observer would find it difficult to distinguish from the original pictures (Goodfellow, 2017). However, they require careful design and tuning to avoid overfitting and training instability (Creswell et al., 2018; Wiatrak et al., 2020). Although they have produced impressive results in image and natural language processing, in some problems they do not outperform SMOTE or ADASYN (Tanaka & Aranha, 2019). As complex deep neural networks, they represent a different approach to the problem of performance prediction, encoding their own features

directly without the need for FPCA, and as such they fall outside the scope of this thesis.

In summary, data augmentation can potentially help improve the accuracy of machine learning models or neural networks, such as at the upper end of the performance range where fewer training data may be available. However, it may require substantial volumes of synthetic data to do so, and even then, the improvements may only be modest at best. It cannot change the fundamentals of the prediction problem, but it can potentially mitigate shortcomings in certain ranges of the outcome variable.

2.7 Conclusion

Jump testing is an important element of an athlete monitoring programme in elite and professional sport. Jump height in the CMJ appears to be the favoured metric rather than peak power based on the small number of surveys available, even though peak power is marginally the more reliable and sensitive measure. The ability to generate power is widely recognised as being vital for sports performance. Such considerations suggest jump height may be favoured for its convenience and intuitive appeal. Indeed, were an alternative method for estimating peak power to become available, one that could be administered easily in the field, then practitioners may welcome it for its practical advantages. However, previous attempts to provide such a solution have failed because they lacked sufficient accuracy, including simple equations based on jump height or Newtonian methods using body-worn sensors. A new approach is needed that will be introduced in this thesis.

Machine learning methods have been employed to estimate VGRF profiles and related discrete measures, but there has been no concerted effort to estimate vertical jump performance. A new approach can be developed for estimating peak power in vertical jumping, drawing on the insights gained from reviewing this literature. Models based on IMUs were not necessarily superior to those using simple accelerometers, as orientation-correcting algorithms were not sufficiently accurate. Neural networks required large data sets but did not produce demonstrably better results than simpler models with smaller samples, as far as VGRF metrics and profiles were concerned. On

the other hand, feature extraction based on FPCA has shown promise in various biomechanical applications, but it has never been employed in machine learning models in anything more than a simple linear regression model. Whilst non-functional PCA is regularly used in machine learning, FPCA's ability to define continuous measures and represent complex patterns more efficiently could be advantageous. It is an approach that has not been tried before in biomechanics.

In the many applications of machine learning, not least in this particular area, it is notable that cross validation methods are not always in evidence (e.g. Table 2-4). Where CV is employed, which is increasingly the case, the K-fold design often does not distinguish between the needs of model selection and model appraisal, even though statisticians have shown the importance of such considerations. Moreover, there were no identifiable examples in biomechanics of using nested cross validation to prevent selection bias. Biomechanics as a discipline is not alone in this regard. Nevertheless, when it comes to developing predictive models that should be expected to perform well on unseen data, a rigorous approach should be followed.

CHAPTER 3. DATA COLLECTION

3.1 Introduction

This chapter concerns the collection of data that will be used throughout this thesis to develop and assess a machine learning model for predicting peak power. The model requires a criterion measure of peak power to serve as the outcome variable that the model will attempt to predict from its various inputs. This will require computing peak power from VGRF data obtained from a force platform using the gold standard method. The jump height can also be determined from the same data to allow comparisons with other studies.

The data set as a whole (VGRF and accelerometer data) should be representative of the intended target population, namely sportsmen and women of a moderate to a high standard. More formally, the data should be drawn from the same distribution (Bishop, 2006; Hastie et al., 2009). Hence, it will be essential to analyse how well the data gathered for this research reflects the jump performances recorded in the literature from sports populations. This can be done by comparing outcome measures such as peak power or jump height, and other aspects, such as fatigue and the effect of arm swing. Checks should also be made for possible signs of bias as that could have a bearing on the models developed later in the thesis.

The size of the data set will also be an important consideration because, all things being equal, a model trained on a larger data set will make smaller errors when applied to new data (Bishop, 2006; Varoquaux, 2018). Previous studies applying machine learning to biomechanical applications have involved relatively few participants, often with each participant performing several trials. For example, the median number of participants was 12 across the studies concerning VGRF-related measures (Table 2-4), while the median trial count was also 12. The single jumping study involved only six participants, but they performed 18 CMJs each (Elvin et al., 2007). Two investigations illustrated a different emphasis between the numbers of participants and the numbers of trials. Pogson et al. (2020) recruited 15 participants who performed 40 running trials, while Derie et al. (2020) sought many more volunteers (93) who carried out

fewer trials of a similar nature (16). In both cases, the total number of training examples was substantial (600 and 1488, respectively).

In contrast, in many biomechanics studies typically only the best of three trials will be used for analysis, such as in vertical jumping where the focus is often on the maximal performance. The studies developing peak power prediction equations generally took this approach (Table 2-1), but they involved large numbers of participants (median 113, range 17–465). So too did the investigations into sensor-based predictions of jump height and peak power, but with fewer participants (12–44; Tables 2-2 and 2-3). The best-of-three convention is usually ignored in ML studies because the overriding aim is to obtain the largest possible data set. The model should be capable of estimating peak power accurately (or any other such performance outcome) because in monitor programmes the jump may not always be performed to the best of the athlete's ability.

The data collection for this thesis will need to recruit a comparatively large number of participants for a biomechanics study. The trials will need to be split between the two different jump types to fulfil the research aims, namely the countermovement jump with and without arm swing, designated CMJ_A and CMJ_{NA}. The data collection can involve men and women as sex was not a significant covariate in the peak power prediction equations (D. Johnson & Bahamonde, 1996; Sayers et al., 1999) or in the sensor-based models of VGRF-related measures (W. Johnson et al., 2017; Meyer et al., 2015; Neugebauer et al., 2012, 2014). Different anatomical positions for the accelerometers should be considered to identify the best location for this application. Therefore, the aims of this chapter are:

- To gather gold-standard VGRF data and obtain criterion measures of jump performance, specifically peak power and jump height, that will serve as the outcome variables in the models developed in this thesis;
- To analyse the performance levels achieved, making comparisons with the literature to establish the validity of the data and its suitability for further analysis in this thesis; and

- To gather accelerometer data from multiple sensors attached to the body whose characteristic features will become predictors in the accelerometer models, although these data will not be examined in this chapter.

3.2 Methods

3.2.1 Participants

A total of 69 physically active men and women volunteered for the study, giving their written informed consent. All but four were involved in sport, classifying themselves as either recreational, club standard or national level athletes (Table 3-1). All participants were required to be free of injury for at least three months prior to data collection to reduce the possibility of compensatory movement patterns influencing the results. For further details on the participants' sporting backgrounds, their self-declared performance level in their chosen sport, as well as other attributes, please refer to Appendix A.1. The Research Ethics and Governance Committee of Swansea University's College of Engineering gave approval for the study.

3.2.2 Data collection

The data collection took place in two stages, with the second introduced to increase the sample size following initial modelling work. The first stage involved 55 participants who performed 16 jumps split equally between countermovement jumps and broad jumps ('jump types'). The requirement for broad jumps arose from a wider research project outside the scope of this thesis. Four jumps were performed for each combination of jump type and arm swing condition (e.g. 4 CMJ_A and 4 CMJ_{NA} for the countermovement jumps). The second stage involved 18 participants who only performed countermovement jumps: 8 CMJ_A and 8 CMJ_{NA}. Four athletes who had taken part previously only performed 4 CMJ_A and 4 CMJ_{NA}, so that over the two stages, they completed a total of 16 countermovement jumps. Although the second data collection made the experimental design unbalanced, this was not considered an issue for the mixed model (Section 3.2.7).

In total, 696 vertical jumps were recorded with equal numbers of CMJ_{NA} and CMJ_{A}. One jump was subsequently discarded from the training/validation set when an issue was discovered with the accelerometer data. The jumps from nine participants were assigned to a 'holdout' test set, with the remainder going to a training/validation data set. The nine were chosen randomly from the 14 volunteers who had taken part in the second data collection, specifically those who had no prior involvement in the first. The models introduced later in this thesis had already been partly developed prior to the second data collection. In this way, the test data was kept entirely separate from the investigation to preserve its independence. In summary, the training/validation set comprised 551 jumps (Table 3-2), and the holdout test set contained 144 (Table 3-3).

3.2.3 Protocol

The warm-up consisted of two minutes of jogging and any other self-directed preparatory exercises the participants preferred. The participants practised each jump type twice, with and without arm swing, using moderate-to-high effort, following a brief demonstration by the investigator. For the CMJ_{NA}, the participants kept their hands on their hips. No advice or feedback was given during the practice or at any other point.

The participants were asked to jump with maximal effort for the data collection. They were given approximately one minute of rest between each jump. The order of jumps was randomised in advance for every participant to minimise potential learning or fatigue effects. The instructor informed the participants what jump they would perform next just before, as the recovery period drew to an end. A jump would be discarded if they landed on the edge of one of the force platforms or could not control the landing. In those cases, the instructor asked them to perform the same jump once the standard rest period was over.

Table 3-1. Combined data set summary (training/validation and testing). Mean ± SD shown for age, body mass and standing height.

	Male	Female	Overall
Number	45	24	69
Age (yrs)	21.2 ± 3.3	22.4 ± 3.6	21.6 ± 3.4
Body Mass (kg)	77.4 ± 12.1	65.2 ± 11.2	73.1 ± 13.1
Standing Height (m)	1.79 ± 0.08	1.65 ± 0.08	1.74 ± 0.10

Table 3-2. Training/validation data set summary. Mean ± SD shown.

	Male	Female	Overall
Number	39	21	60
Age (yrs)	21.3 ± 3.3	22.1 ± 3.7	21.6 ± 3.5
Body Mass (kg)	77.7 ± 12.8	63.9 ± 10.9	72.9 ± 13.7
Standing Height (m)	1.80 ± 0.08	1.65 ± 0.08	1.74 ± 0.11

Table 3-3. Testing data set summary. Mean ± SD shown.

	Male	Female	Overall
Number	6	3	9
Age (yrs)	20.5 ± 2.9	24.7 ± 3.7	21.9 ± 3.2
Body Mass (kg)	75.2 ± 6.3	73.9 ± 11.9	74.8 ± 7.8
Standing Height (m)	1.75 ± 0.07	1.68 ± 0.01	1.73 ± 0.06

3.2.4 Force Platforms

The jumps were performed from two portable force platforms (9260AA, Kistler, Winterthur, Switzerland), 400 mm × 600 mm, placed on Everoll flooring, a composite rubberised surface (Regupol BSW, Bad Berleburg, Germany). The force platforms were placed adjacent to one another with only a slight gap between them so that the participants could stand with one foot on each plate. Two force platforms were employed to analyse bilateral asymmetries as part of the same wider project mentioned above that was outside the scope of this thesis. The force platforms were connected to a 16-bit analogue-to-digital converter (5691A, Kistler) with a sampling rate of 1000 Hz. The Nexus v2.5 software (Vicon, Oxford, UK) synchronised the ground reaction force data with the accelerometer data (Section 3.2.5). The investigator used the VGRF trace presented in the Nexus software to determine when the participant was standing still, at which point he instructed them to “jump when ready.” The data

recording continued until they had regained a stationary upright position after landing. The force platforms were calibrated on each day of testing and zeroed immediately before the participant stepped onto them. (Please refer to Appendix A.2 for further details of the calibration procedure that ensured both force platforms registered the same VGRF for the same weight.)

3.2.5 Wearable sensors

Delsys Trigno sensors (Delsys Inc., Natick, MA, USA) were used in the present study as they had the best fit for the research requirements. The Delsys sensors had high validity and reliability, and they offered automatic synchronisation with other accelerometers and with force platforms. Appendix B discusses the steps taken to select the most appropriate sensors from the laboratory's inventory. The Delsys sensors (dimensions 27 × 37 × 15 mm, mass 14.7 g) had onboard triaxial accelerometers with a stated maximum range of ± 9 g. However, the calibrated analogue signals recordings indicated that the devices could quantify inertial accelerations up to ~10.5 g. This range was sufficient to record the inertial accelerations of the trunk and lower limbs generated during take-off and landing.

The sensors continuously transmitted their measurements to a Delsys Trigno base station, which relayed them via a multi-channel, analogue connector to a Vicon MX-Giganet box. The Gigabit box was connected to a Windows PC via a 1 GHz Ethernet cable and controlled data communications from multiple devices in the Vicon Vantage system. The Vicon Nexus software imported the continuous accelerometer data streams, sampled at 250 Hz, synchronising them with the VGRF data. The Delsys system's automatic synchronisation was a significant advantage over other products assessed in Appendix B.3.

The sensors were attached directly to the participant's skin using double-sided surgical tape at the following anatomical locations (Figure 3-2): the lower back over the L4 vertebrae; the upper back over the C7 vertebrae; and the lower anterior medial aspect of the tibias located by taking 40% of the length of the shank, measured vertically from the medial malleolus. The shank length was measured using callipers from the medial malleolus to the medial epicondyle located by palpation with the participant standing.

Tiger Tape (Physique, Havant, UK), an elastic adhesive bandage, held the sensors firmly in position. Compressive attachments are recommended to minimise sensor oscillations due to soft-tissue movement (Forner-Cordero et al., 2008; Ziegert & Lewis, 1979).

The four sensors were calibrated, one at a time, by placing them in six stationary, orthogonal orientations. One of the onboard accelerometers was aligned with gravity in each position, while the other two were perpendicular. One of the portable force platforms provided a convenient set of orthogonal surfaces for the purpose, as the platforms were perfectly level, having been checked using a spirit level (Figure 3-1). Taking each sensor in turn, they were first placed flat on the top surface of the force platform (position 1, where $X = 0,\ Y = 0,\ Z = 1$). The reverse orientation with the sensor upside down was not part of the calibration because it had a convex top. The sensor was then held in a vertical position at a corner of the force platform (positions 2 and 3), using the corner's right angles to ensure correct alignment. The sensor was held in four static orientations: 3'o clock ($X = 0,\ Y = 1,\ Z = 0$), 6'o clock ($X = 1,\ Y = 0,\ Z = 0$), 9'o clock ($X = 0,\ Y = -1,\ Z = 0$) and 12'o clock ($X = -1,\ Y = 0,\ Z = 0$). For all five orientations, an average was taken over a suitable static period in the accelerometer signal.

Calibration equations were determined for each sensor's axis for each day of data collection using a linear best fit of the criterion and averaged measurement accelerations. The best fits achieved $r^2 = 1.000$ to three decimal places in almost all cases ($4 \text{ sensors} \times 3 \text{ axes} \times 24 \text{ days of testing} = 288$ cases), or otherwise, it was $r^2 = 0.999$. The standard error in the calibration equations' slopes averaged over all sensors, axes and days of testing was 0.011 $g \cdot V^{-1}$, equivalent to 0.33%. Similarly, the standard error in the calibration equation intercept was 0.006 g.

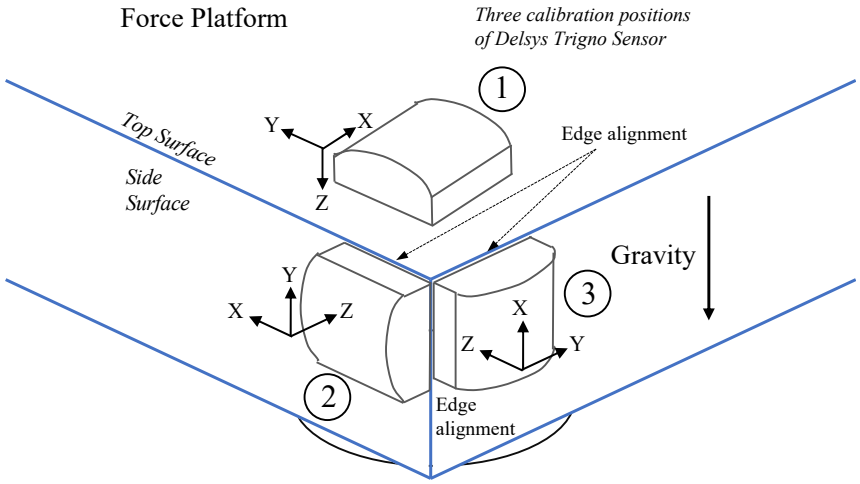
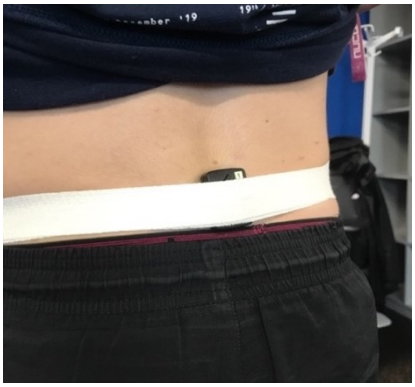
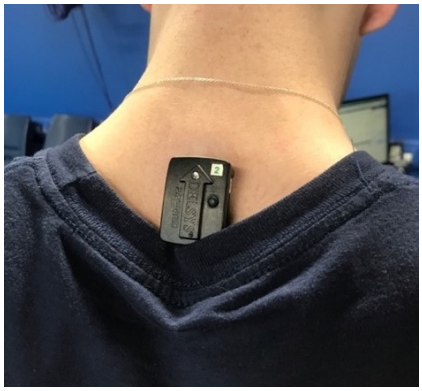


Figure 3-1. Sensor calibration setup employing the force platform's orthogonal surfaces for correct alignment with respect to gravity. The criterion values for the accelerometers were: for position (1), $X = 0$, $Y = 0$, $Z = 1$; for position (2), $X = 0$, $Y = -1$, $Z = 0$; and for position (3), $X = -1$, $Y = 0$, $Z = 0$. Positions (2) and (3) were also reversed so that $Y = 1$ and $X = 1$, respectively, with the other axial directions equal to zero.

(A)



(B)



(C)

(D)

Figure 3-2. Sensor attachments. (A) Lower back (L4); (B) Upper back (C7); (C) Left shank; (D) Right shank (tibial anterior medial aspect for each leg). The sensors were fixed to the skin using hypoallergenic double-sized tape, reinforced using Tiger Tape to hold the sensor in place firmly. This technique was impractical for the upper back sensor.

3.2.6 Calculation of jump performance

All data processing was carried out by custom scripts in Matlab. The code calibrated the VGRF data from each force platform and then added them together to give a total VGRF. It did not apply any filtering in line with published recommendations when processing jump GRF data (Owen et al., 2014; Street et al., 2001). All calculations of jump performance were based on the vertical component of the GRF alone.

According to Newton's Laws of Motion, the vertical velocity, $v(t)$, is given by the time integral of the net VGRF, $F_{\mathrm{Net}}(t)$, with bodyweight subtracted:

$$v(t) = \frac{1}{m} \int F_{\mathrm{Net}}(t)dt \tag{3.1}$$

where m is the body mass. The bodyweight (BW) was determined initially for each jump by taking the mean VGRF over the first second of recording when the participant was standing still. The BW estimates varied very slightly across all jumps for the same individual, where the standard error (SE) in BW across all participants was $0.0021 \pm 0.0018\%$.

Since even small errors in BW estimates can lead to large jump height errors (Street et al., 2001), an algorithm was devised to improve the BW estimates. The algorithm searched for a period when the VGRF had the smallest standard error, from the start subject to a minimum duration of 250 ms and a maximum of 2000 ms. This algorithm reduced the SE in the BW estimate to $0.0014 \pm 0.0005\%$. This marked a modest reduction of one third, but it eliminated outliers which had been a drawback of the standard method. Manual checks were made by inspecting the vertical displacement curves from the start of the recording to check there was no appreciable drift upwards or downwards (< 0.001 m) before the jump began. This observation was verified by overlaying the accelerometer signals, which also indicated when the movement started. Small manual corrections to the BW were made, if necessary, typically of the order of 0.1 or 0.2 N.

The vertical displacement, $s(t)$, is the time integral of the velocity:

$$s(t) = \int v(t)dt \tag{3.2}$$

From equation (3.2), jump height was defined as the maximum value of $s(t)$, the work-done definition, corresponding to the total energy required to raise the body's CM to this height from the standing position.

The instantaneous power as a function of time, $P(t)$, is given by:

$$P(t) = F(t)v(t) \tag{3.3}$$

From equation (3.3), the peak power is the maximum value of $P(t)$, representing the maximal rate of production of external mechanical energy. For the analysis in this and subsequent chapters, the instantaneous power was normalised to the participant's BW to obtain relative power. Hereafter, unless otherwise stated, the term peak power refers to the peak relative power.

All numerical integrations were performed using the trapezoidal rule with a time interval of 1 ms, from jump initiation to take-off, defined as VGRF < 10 N. The jump initiation point was determined using an algorithm adapted from the method proposed by Owen et al. (2014), which starts by detecting the first sign of movement where the VGRF diverged from BW by more than 5 × SD. The algorithm searched backwards from this point to find an earlier point where the jump began, using stricter criteria, unlike the fixed 30 ms offset employed by Owen et al. (2014). This more adaptable approach was necessary to handle the greater variety of VGRF patterns seen across 500+ jumps. Further details of the algorithm are presented in Appendix B.3.

3.2.7 Statistical Analysis

Mixed statistical models were developed with random intercepts to predict peak power and jump height using SAS Studio 3.8 (SAS Institute Inc., Cary, NC, USA). These outcome variables were log-transformed to enable percentage comparisons over a wide range of performances. The mixed models do not require independent observations, unlike traditional regression models.

The mixed models were implemented using the MIXED procedure based on restricted maximum likelihood with the arm swing condition and sex as the fixed effects. Sex was included to facilitate comparisons with other studies. The possible impact of

fatigue was accounted for by the jump trial number as a fixed effect. Since the randomisation of the jump order could also produce potential biases – a string of repeated jumps of the same type or the switching between jump types – additional fixed effects were included as dummy variables. A jump repetition variable equalled one if the current jump was the same as the previous one, otherwise the variable was zero. A jump switching variable equalled one if there was a switch from vertical to horizontal jumps or vice versa (in the first data collection), otherwise it was zero. These covariates were retained in the model if their corresponding type-III F-statistic was significant (α = 0.01). The overall model fit was checked using the Bayesian Information Criterion (BIC), a conservative measure suitable for large datasets that penalises complexity.

The estimates reported for the fixed effects on jump height and peak power were based on the model that included only the significant covariates. Those estimates were the least-squares means differences (also known as the marginal means differences), following the reverse log transformation. The results were interpreted using the standard effect size thresholds based on Cohen's *d* of 0.2 (small), 0.5 (medium), 0.8 (large) and 1.2 (very large) (Cohen, 1992). Where $d < 0.2$, the effect size is regarded as trivial. The confidence limits were set at 90%.

The covariance matrix was specified with only the model intercept as the sole random effect, grouped by the participant ID. The covariances were unbounded. The variance in the model intercept accounted for the average variation in performance between individuals. The Smallest Worthwhile Change (SWC) was taken to be one-fifth of this value (W. G. Hopkins, 2004). The model residual represented the typical variation in an individual's performance from jump to jump (W. Hopkins, 2016). The model was re-fitted for the jump-type subsets so that the random effects could be categorised for CMJ_{NA} and CMJ_{A}.

3.3 Results

Higher performance levels were achieved, on average, in jumps with arm swing compared to those without, and in men compared to women (Tables 3-1, 3-2, 3-3 and

3-4). There was a close similarity between the training/validation and testing data sets in terms of mean, standard deviation and the 10^{th} and 90^{th} percentiles. The distributions show skew towards higher peak power outputs in both arm swing conditions and skew towards lower jump heights (Figure 3-3). The distributions encompass a broader range of peak powers and jump heights, as indicated by a negative kurtosis (Figure 3-3).

The best jump heights were achieved by some volleyball players and swimmers in both arm swing conditions (> 0.55 m for CMJ_{NA} and > 0.70 m for CMJ_A), but their corresponding peak power output was not at the same top end of performance (most were < 60 $W \cdot kg^{-1}$). The volleyball players' jump performances were notable for having a bimodal distribution, unlike those from other sports. Football players had the widest spread of performances, while rugby players, netball and hockey players had narrower, localised ranges. Note that the results presented in this chapter are based on the VGRF data. The accelerometer data will be assessed later in the thesis, starting in Chapter 5, which will develop the first accelerometer models.

Table 3-4. Peak power outputs for the training/validation data set.

Peak Power ($W \cdot kg^{-1}$)		Mean ± SD	10th – 90th Percentile	Min, Max
Overall	**CMJ_{NA}**	**45.0 ± 7.3**	**35.2 – 54.1**	**27.2, 63.6**
	CMJ_A	**51.5 ± 8.6**	**39.7 – 62.1**	**28.1, 72.5**
Males	CMJ_{NA}	48.4 ± 5.4	41.4 – 54.9	36.6, 63.6
	CMJ_A	55.6 ± 6.3	47.5 – 63.3	38.2, 72.5
Females	CMJ_{NA}	38.2 ± 5.7	31.3 – 43.6	27.2, 55.9
	CMJ_A	43.3 ± 6.7	35.7 – 50.9	28.1, 62.0

Table 3-5. Peak power outputs for the testing data set.

Peak Power ($W \cdot kg^{-1}$)		Mean ± SD	10th – 90th Percentile	Min, Max
Overall	**CMJ_{NA}**	**47.6 ± 8.1**	**33.0 – 55.9**	**29.4, 59.0**
	CMJ_A	**53.4 ± 10.0**	**34.9 – 63.9**	**31.6, 67.0**
Males	CMJ_{NA}	51.5 ± 4.1	45.8 – 56.5	44.8, 59.0
	CMJ_A	58.4 ± 4.8	53.0 – 66.1	45.7, 67.0
Females	CMJ_{NA}	39.7 ± 8.3	30.9 – 51.2	29.4, 52.9
	CMJ_A	43.3 ± 9.9	31.7 – 57.8	31.6, 58.5

Table 3-6. Jump heights for the training/validation data set.

Jump Height (m)		Mean ± SD	10th – 90th Percentile	Min, Max
Overall	**CMJ_{NA}**	**0.400 ± 0.085**	**0.292 – 0.501**	**0.223, 0.636**
	CMJ_A	**0.479 ± 0.099**	**0.364 – 0.604**	**0.244, 0.765**
Males	CMJ_{NA}	0.441 ± 0.067	0.361 – 0.534	0.227, 0.636
	CMJ_A	0.525 ± 0.083	0.415 – 0.631	0.261, 0.765
Females	CMJ_{NA}	0.317 ± 0.048	0.274 – 0.367	0.223, 0.483
	CMJ_A	0.385 ± 0.050	0.343 – 0.444	0.244, 0.531

Table 3-7. Jump heights for the testing data set.

Jump Height (m)		Mean ± SD	10th – 90th Percentile	Min, Max
Overall	**CMJ_{NA}**	**0.422 ± 0.088**	**0.280 – 0.543**	**0.246, 0.589**
	CMJ_A	**0.497 ± 0.122**	**0.334 – 0.697**	**0.279, 0.738**
Males	CMJ_{NA}	0.461 ± 0.063	0.394 – 0.569	0.356, 0.589
	CMJ_A	0.546 ± 0.101	0.426 – 0.726	0.363, 0.738
Females	CMJ_{NA}	0.343 ± 0.078	0.249 – 0.448	0.246, 0.451
	CMJ_A	0.400 ± 0.100	0.293 – 0.547	0.279, 0.562

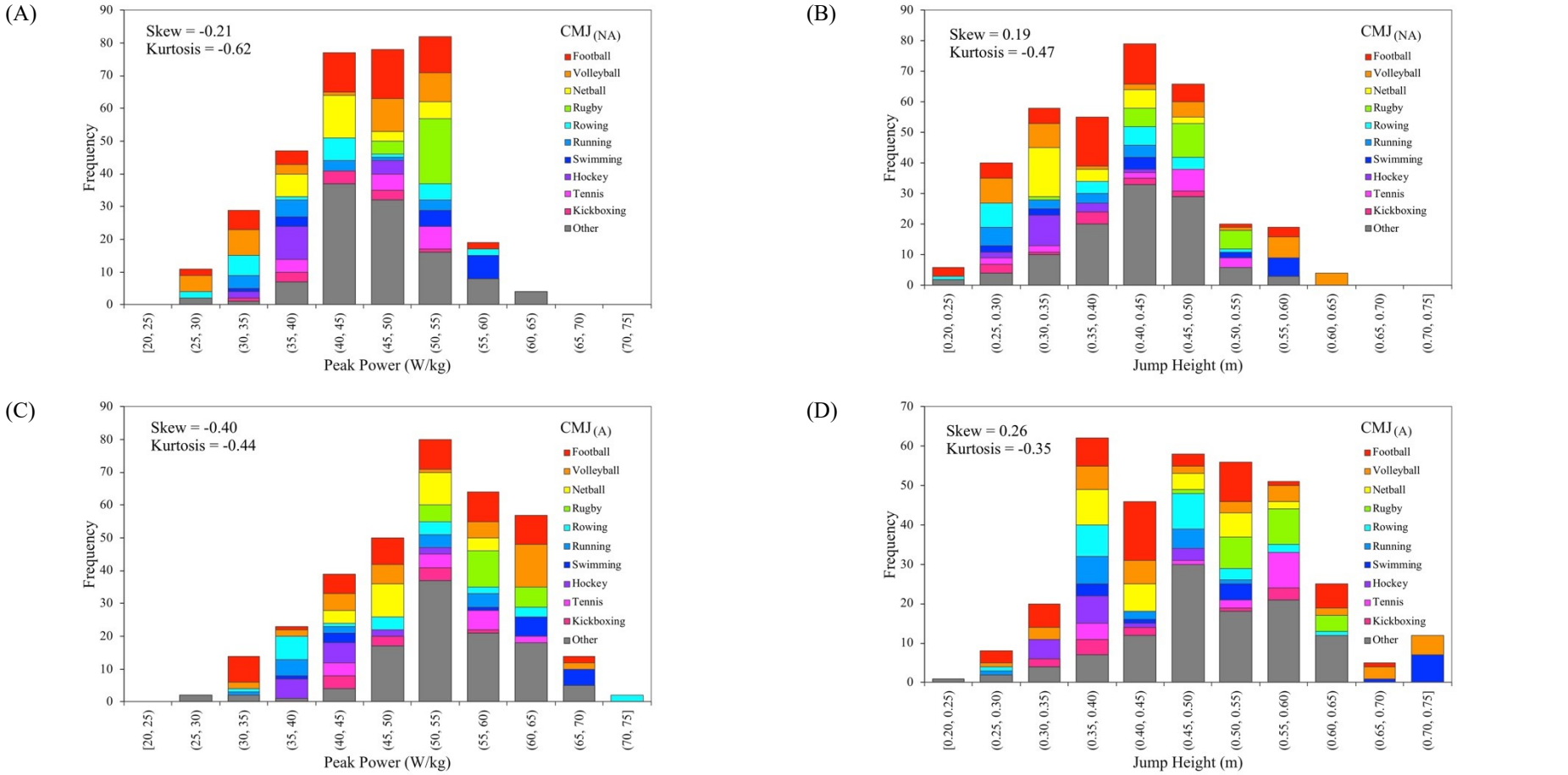


Figure 3-3. Distribution of jump performances stratified by primary sport, covering both sexes for each jump performance measure. (A) Distribution of peak power in the CMJ_{NA} showing a skew towards higher values, notably from rugby players (green). (B) Distribution of jump height in the CMJ_{NA} with a long upper tail that includes jumps from volleyball players (orange) and swimmers (deep blue). (C) Distribution of peak power in the CMJ_{A} also with a skew towards higher performances, including jumps from participants in several sports. (D) Distribution of jump height in the CMJ_{A} with a broader range of performances, with notable outliers at the top end featuring jumps from volleyball players and swimmers (orange and blue). Individual sports are identified where at least three participants stated this was their primary sport. All other sports are classified as 'other'.

The mixed models revealed that arm swing, sex, and jump trial significantly affected peak power and jump height (Table 3-8). The possible confounding factors related to the jump order were trivial and non-significant. Arm swing had a similar effect on peak power and jump height, centred around the boundary between the categories of 'medium' and 'large' effects. Sex had only a small effect on peak power but a medium effect on jump height. The fatigue effect, indicated by the jump trial number, was greater on peak power than jump height, a 'small' versus a 'trivial' effect, respectively. With the non-significant effects removed, the arm swing effects on peak power and jump height were approximately 13.7% and 19.2%, respectively (Table 3-9). In comparison, these percentage differences were smaller than those between men and women.

Table 3-8. Effect sizes for all fixed effects for the full model (Type III tests).

	Peak Power	Jump Height
Arm Swing	0.806 ** (L)	0.795 ** (M)
Sex	0.329 ** (S)	0.735 ** (M)
Jump Trial Number	0.266 ** (S)	0.106 * (T)
Jump Repetition	0.020 (T)	0.014 (T)
Jump Type Switch	0.034 (T)	0.012 (T)

** $p < 0.0001$. Effect sizes: L = Large; M = Moderate; S = Small; T = Trivial.

Performance declined over multiple jumps, based on the jump trial number. The peak power dropped by 0.30% on average with each jump, while the jump height fell by 0.16%. Over 16 jumps, this added up to a decline in peak power of approximately 4.8% and a fall in jump height of 2.6%. When the jump trial number was treated alternatively as a random effect, the model fit (BIC) improved, and the residual error for peak power predictions reduced slightly from 5.1% to 4.7%. The error declined from 7.3% to 6.8% when the jump trial number was a random effect in the jump height model. Making arm swing a random effect did not improve the model.

Table 3-9. Fixed effects on performance measures for the compact model

	Peak Power	Jump Height
Arm Swing Effect	13.7 ± 0.7% **	19.2 ± 1.1% **
Sex Effect (M vs. F)	22.5 ± 3.7% **	27.5 ± 4.3% **
Trial Effect (per jump)	-0.30 ± 0.07% **	-0.16 ± 0.10% *

Difference of Least Squared Means shown with 90% confidence intervals.
* $p < 0.001$; ** $p < 0.0001$.

The typical variation in an individual's performance was lower in the CMJ_{NA} than the CMJ_A for peak power and jump height (Table 3-10). The smallest worthwhile change, calculated from the between-individuals variance, was 2.5% for peak power and 3.0–3.1% for jump height.

Table 3-10. Random effects for jump-type specific compact models.

	Jump Type	Peak Power	Jump Height
Within-Individual SD	CMJ_{NA}	4.0 ± 0.3% **	6.5 ± 0.5% **
	CMJ_A	5.3 ± 0.4% **	7.3 ± 0.4% **
Smallest Worthwhile Change	CMJ_{NA}	2.5 ± 0.4% **	3.0 ± 0.5% **
	CMJ_A	2.5 ± 0.4% **	3.1 ± 0.5% **

90% confidence intervals shown.
* $p < 0.001$; ** $p < 0.0001$.

3.4 Discussion

This chapter described the collection of VGRF and accelerometer data and the calculation of the criterion performance measures, the so-called 'ground truth' in machine learning that will be used in models developed later in this thesis. The accelerometer data will be used from Chapter 5 onwards. Since the VGRF and accelerometer data will provide the inputs for subsequent extensive analysis throughout this thesis, it is essential to assess whether the results are representative by comparing them to those reported in the literature.

3.4.1 Comparisons of performance levels

Studies reporting peak power for the CMJ_{NA} cover a wide range of performance levels for elite male athletes, ranging from 45.4 $W \cdot kg^{-1}$ for 'elite rugby players' (Marrier et al., 2017) to 54 $W \cdot kg^{-1}$ for 'elite Australian rules football players' (Cormack, Newton, & McGuigan, 2008), and 65.1 $W \cdot kg^{-1}$ for 'male, college-level team-sport athletes' (Gathercole et al., 2015). Compared to those standards, the mean power output of 48.4 $W \cdot kg^{-1}$ for males in the current study was at the lower end of this range. The rugby players from the present study achieved higher performances in the 50-55 $W \cdot kg^{-1}$ range, which compares well with the studies by Marrier et al. (2017) and Cormack, Newton & McGuigan et al. (2008). The female participants from the current study produced peak power outputs of 38.2 $W \cdot kg^{-1}$ on average in $CMJ_{NA,}$ which is in-between the 34.8 $W \cdot kg^{-1}$ achieved by 'untrained female college students' who played sports recreationally (Makaruk et al., 2011) and the 43.4 $W \cdot kg^{-1}$ recorded by female NCAA volleyball players (Newton et al., 2006). In one of the few studies looking at peak power in the CMJ_A, 'physically active' men, who were not resistance-trained but played Australian-rules football recreationally, achieved peak power outputs of 58.9 $W \cdot kg^{-1}$ on average, which compares with 55.6 $W \cdot kg^{-1}$ for the males in the current study. These comparisons were based on group means, but it is also worth noting that the top performers in the present study achieved peak power outputs comparable to the elite athletes in the studies by Cormack, Newton & McGuigan et al. (2008) and Gathercole et al. (2015). In conclusion, the peak power outputs recorded in the present study are comparable to those in the literature, giving confidence that the current data set represents moderate to high performers.

Several studies used the work-done definition of jump height, the method that was used to calculate the results in this chapter. The mean jump heights for CMJ_{NA} were 0.54 m for 'trained male volleyball players' (Bobbert et al., 1987a), 0.52 m for 'physically active male college students' (Aragón-Vargas & Gross, 1997), and 0.495 m for a mixed cohort of 'healthy young adults' (31M: 9F) (Buckthorpe et al., 2012). In general, these performances were higher than those in the present study (0.441 m for males, 0.400 m across both sexes). However, the four male volleyball

players from this cohort achieved a jump height of 0.527 m on average, comparable to the volleyball players in Bobbert's (1987) study above.

Other investigators have used the flight-time definition of jump height, which ignores the rise in the body's CM before take-off. Consequently, the reported jump heights appear to be noticeably lower than those cited above. The jump heights in the present study were recalculated using the flight-time definition to allow comparison. Accordingly, the jump heights were 0.296 m and 0.339 m for the CMJ_{NA} and CMJ_{A}, respectively, averaged across both sexes. Comparable figures can be seen in a mixed group of physical education students (73M:33F), including ten elite athletes from volleyball and weightlifting, who attained jump heights of 0.29 m and 0.35 m, respectively (Oddsson, 1987). Similar results were seen in 'moderately to highly trained' men and women (0.30 m) (Dowling & Vamos, 1993).

Comparisons can also be made with studies involving either exclusively men or women. Jump heights were reported of 0.35 m for 'healthy college-age men' (G. Markovic et al., 2004), 0.401 m for elite junior-level basketball players (Apostolidis et al., 2004), based on flight time, compared to 0.335 m in the current study. Using the same measure, professional female volleyball players registered jump heights of 0.287 m (González-Ravé et al., 2011) compared to 0.217 m in the present study. Another investigation using the work-done definition reported jump heights of 0.400 m for female volleyball players in the NCAA (Newton et al., 2006) versus 0.317 m, as presented above.

These comparisons reveal the present investigation's performance levels were similar to those reported in the male-only and unisex studies but lower than in other female-only cohorts. The female participants in the present study were from various sports and backgrounds, with only a few considering themselves to be of national standard. The female volleyball players were comparative novices with only 1-2 years' experience, lacking the proficiency of their more experienced male counterparts. In contrast, the cited research involved professional or NCAA-level volleyball players. This example serves to underline the difficulties of making comparisons between studies, often with small samples. Setting aside elite-level comparisons, the women in the current investigation jumped similar heights to their female counterparts in the

literature. The jump height results support the preliminary conclusion that the present study's jump performances are valid and representative.

3.4.2 Arm swing effects

Another way to assess the data's validity is to see how the arm swing effect compares with other studies. The fixed arm swing effect on peak power production was smaller than it was on jump height (13.7 ± 0.7% vs. 19.2 ± 1.1%), suggesting the main contribution to neuromuscular power comes from the lower extremity. Comparisons can only be made with the literature concerning jump height. The arm swing effect above was in close agreement with most other studies, ranging from 18.5% in 'healthy males' (Domire and Challis, 2010), 19.3% in 'athletic adult males' (Lees et al., 2004) to 20.7% in 'physical education students' (Oddsson, 1987). Given how similar the figures are between studies, the effect of arm swing on jump height appears to be relatively constant despite differences of strength or skill. This finding was supported by a longitudinal study of volleyball players over two years (Borràs et al., 2011). There was only a marginal change in the arm swing effect from 20.2% to 20.7%, even though jump height increased in response to training (Borràs et al., 2011). In other words, arm swing helps raise jump height by approximately the same proportion in physically active individuals. This observation is reflected in the finding that the model fit was better when arm swing was treated as a fixed effect rather than a random effect. Hence, individual variations in arm swing technique may not have a significant bearing on the models.

3.4.3 Fatigue

The trial jump number, a proxy for fatigue, exerted a more sizable influence on peak power than on jump height. Over 16 jumps, peak power declined on average by 4.8% compared to a drop in jump height of 2.6%. Gathercole et al. (2015) reported that the jump performance variables, peak and mean power were among the variables that had not returned to baseline levels 24 hours after heavy exercise. Several other studies have found peak power and flight time (a proxy for jump height) to be valid indicators of impaired neuromuscular function in professional players in team sports (Andersson et

al., 2008; Claudino et al., 2017; Johnston et al., 2013; McLean et al., 2010). The fact that jump performances tended to decline across the test session is not a concern for this investigation and the subsequent development of an accelerometer model. Given the purpose of jumping testing is for monitoring athlete training status, some degree of participant fatigue would be desirable.

3.4.4 Random effects

The random effects divided the fixed model's unexplained variance into the variance within and between individuals. The within-individuals variance represents the consistency of performances from jump to jump on average across the cohort. It was 4.0% for peak power in the CMJ_{NA}, which for the mean performance would be equivalent to 1.8 $W \cdot kg^{-1}$. In absolute terms, this figure agrees with the CV figure averaged over four studies (1.75 $W \cdot kg^{-1}$) and falls within the range of percentages reported, 2.3%–5.2% (Cormack, Newton, & McGuigan, 2008; Gathercole et al., 2015; McLellan et al., 2011; Taylor et al., 2010). It is also notable that peak power exhibited lower variance within and between individuals than it did for jump height, as other studies have reported (Section 2.2.3). The smallest worthwhile change will be a factor to consider when reviewing the practical usefulness of the final model in the conclusion to this thesis (Section 9.4.1).

3.4.5 Data set composition

The results showed that sex had a significant effect on both performance measures. Men tended to jump higher or generate more power than women, as expected. However, sex was not a significant covariate in models predicting performance metrics, where it has been considered when using jump height or inertial accelerations as inputs (Sections 2.3.2 & 3.1). In sprinting, certain kinematic parameters were more strongly influenced by the sprinters' performance level than by their sex (Ciacci et al., 2017). More generally, a review into the differences between men and women in athletic movements found little support for sex-specific, lower limb kinematic patterns for landing and hopping, nor sex differences in muscle activation (Bruton et al., 2013). Whilst there are undoubted group differences in performance between men and

women, those differences do not appear to be due to sex, *per se*. In developing models based on inertial accelerations, the focus will be on how those performances were achieved rather than the outcomes themselves. Therefore, it would seem reasonable to assume that models for predicting peak power in vertical jumping can be developed using data from men and women.

3.4.6 Possible confounding effects

The mixed models proved to be powerful analytical tools, allowing checks for potential confounding factors. The participants performed broad jumps and countermovement jumps in the first data collection, but the jump switch effect was close to zero. Thus, the possibility that the horizontal jumps influenced the vertical jumps can be dismissed. The random order in the jumps (of whichever type and arm condition) could have been a confounding factor. Four participants repeated the same jump three times, two repeated the same jump four times, one did five repeats, and another participant repeated the same jump six times in a row. While such repetition may produce a learning effect, such consecutive series are rare. Hence, the overall effect was trivial, as the jump repeat effect showed.

3.4.7 Conclusions

This chapter described the collection of VGRF and accelerometer data that will be used throughout this thesis. The performance measures computed from the VGRF data, peak power and jump height, were the 'ground truth' and will serve as the outcome variables in subsequent models. Many of the participants were competent jumpers, with several coming from a background in volleyball and basketball. They produced jump performances that were comparable with those reported in the literature for moderate to high performers. The effects of arm swing and fatigue were also consistent with previous research, indicating these performances were representative of jump testing programmes. These comparisons also provided an indirect check on the validity of the methods described in this chapter. Therefore, this data can confidently be used for model development and further analysis in this thesis, starting

in the next chapter with models based on VGRF data. Models based on the accelerometer data will be developed from Chapter 5 onwards.

CHAPTER 4. VERTICAL GROUND REACTION FORCE MODELS

4.1 Introduction

In this chapter, models based on the VGRF data are developed to estimate peak power and jump height from features extracted using FPCA. It will address the first research question, which asked *how well does an FPCA-type model perform when predicting peak power and jump height in the CMJ?* The first models using FPCA to predict a performance outcome were developed by Richter et al. (2014a), who estimated CMJ jump height from VGRF data. As discussed in Section 2.5.4, the FPCA model had a fit of $r^2 = 0.79$, but the corresponding ACP model in that study achieved $r^2 = 0.98$. In comparison, Moudy et al. (2018) could only achieve a more modest fit with ACP of $r^2 = 0.86$ with the benefit of curve registration. No previous study has used an FPCA-based model to predict peak power in vertical jumping. If such a model proves effective with gold standard VGRF data, then this approach can be applied to accelerometer data with confidence that FPCA-based modelling is not a limiting factor.

FPCA is dependent on the curves being in close alignment with one another (Section 2.5.3). The peaks and troughs from different curves should broadly coincide so differences in their magnitudes can be recognised, otherwise, cross-sectional measures of amplitude variance, the so-called modes of variation, will be diminished (Ramsay & Silverman, 2005). However, there will be some misalignment due to different movement strategies, varying jump execution times and natural movement variability. Richter et al. (2014a) and Moudy et al. (2018) time-normalised the VGRF data to a standard duration, as have other investigators (e.g. Kennedy and Drake, 2018). There was already a high degree alignment in Richter's case because the data were restricted to the propulsion phase, which has a duration that is the most consistent of all the jump phases (J. McMahon et al., 2017; Oddsson, 1987; van Ingen Schenau, 1989).

Moudy et al. (2018) utilised registration to bring the curves into closer alignment, reporting that a single landmark was best, located at the start of the propulsion phase.

A similar approach can be taken for the present investigation in this chapter to determine which landmarks are appropriate for a peak power model. However, it would also be worth considering padding out the VGRF curves to a standard length rather than time-normalising them in order to preserve the time domain (Preatoni et al., 2013). Epifanio et al. (2008) found that an FPCA-based classification model was more accurate when padding rather than time-normalisation was used. With padding, it would be preferable to align the VGRF curves at take-off, as is often done in jumping studies (e.g. Bobbert & van Ingen Schenau, 1988; Dowling & Vamos, 1993). Curve registration may still be required as misalignment becomes more apparent when inspecting the curves further away from take-off.

Therefore, the aim of this chapter is to:

- Determine the efficacy of FPCA-based models for predicting peak power and jump height based on gold-standard VGRF data; and
- Evaluate the techniques for achieving curve alignment by comparing padding versus time-normalisation and various landmarks for curve registration.

4.2 Methods

4.2.1 Data preprocessing

The models developed in this chapter will be based on the VGRF training/validation data set, sampled every 1 ms (i.e. 1000 Hz). Each time series began at the jump initiation point (defined as when the VGRF first deviated from BW by $> 5 \times$ SD, subject to a time adjustment – Appendix A.3) and ended at take-off (VGRF < 10 N). The distribution of jump execution times was strongly skewed towards shorter times with a median of 1.283 s (Figure 4-1), with 95% of the times falling within the interval [0.954 s, 2.269 s].

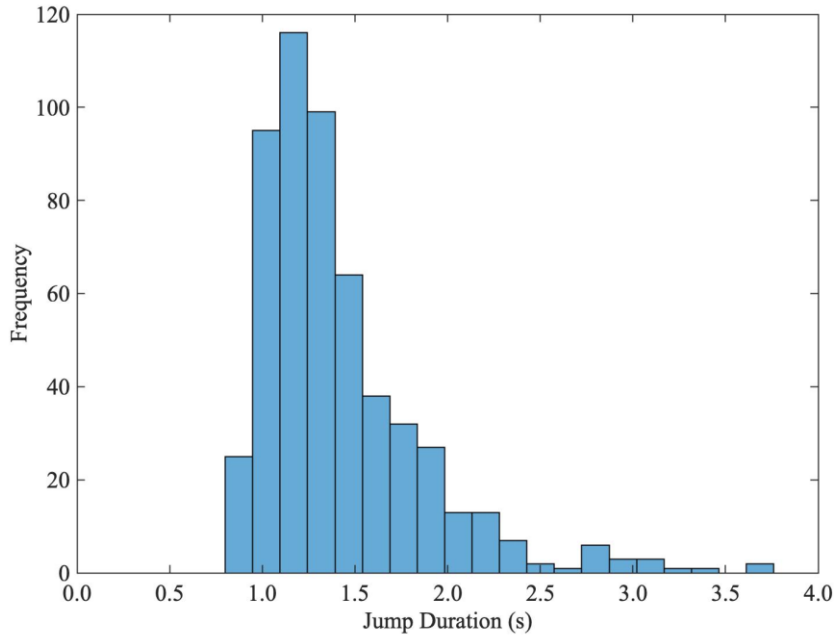


Figure 4-1. Distribution of jump execution times measured from the jump initiation point to take-off.

Two methods were used to standardise the curve duration. The first method was linear time normalisation (LTN), where each time series is re-sampled to a fixed number of points using cubic interpolation. The LTN window size was set to 1280 points, the rounded median number of points, so adjustments to the timeframe were kept to a minimum. The second method padded out the time series (PAD) by inserting a series of 1's at the beginning (bodyweight for the normalised VGRF data), equivalent to the participant standing still before the jump. The PAD curves' standard length was 2250 points, equivalent to 2.25 s. Jumps longer than this were excluded in both methods (28 jumps or 5.1% of the total) to make for a fairer comparison by not disadvantaging LTN unduly since some jumps took considerably longer than others because some participants rehearsed their arm swing. Consequently, the data set was reduced to 520 jumps, comprising 254 CMJ_A and 266 CMJ_{NA}. Although the LTN and PAD data sets had different lengths (1280 points for LTN and 2250 points for PAD), they had roughly the same density of points per unit time. For LTN, this density varied slightly between jumps, whereas for PAD it was constant.

4.2.2 Functional smoothing

The LTN and PAD data sets were converted into smooth continuous functions defined by a linear combination of b-spline basis functions. The number of basis functions, K, was chosen so the overlapping b-spline functions each spanned five data points ($K = 256$ for LLN and $K = 450$ for PAD). The b-splines were 5^{th}-order to ensure the function had a smooth curvature with no discontinuities. Generalised cross-validation was used to determine the basis function density and the 3^{rd}-order roughness penalty ($\lambda = 10^{-1}$). The roughness penalty was high because the time domain was scaled as that one time unit was equivalent to one millisecond in reality. This scale was necessary to avoid numerical errors later in the analysis (Ramsay & Silverman, 2005).

4.2.3 Registration

The smoothed VGRF curves were brought into closer alignment with one another using landmark registration to warp each curve's time domain appropriately (Ramsay & Li, 1998). The registration criterion was to minimise the mean square error between each curve and the cross-sectional mean curve computed before registration. The warped time domains were defined using a smooth, monotonic continuous function comprising ten 1^{st}-order, b-spline basis functions ($\lambda = 10^{1}$). The same four landmarks used by Moudy et al. (2018), which represent changes of phase or direction in the jump, were considered (all 16 combinations) along with the baseline case of no registration: VGRF minimum (designated 'L1'), power minimum ('L2'), the start of the propulsion phase ('L3') and peak power ('L4'). The landmarks could be identified unambiguously using standard methods to locate axis-crossings and turning points. When applied to both LTN and PAD data sets, registration produced 32 data sets altogether for evaluation in the subsequent models (Section 4.2.5).

4.2.4 Feature extraction

The FPCA procedure defined the FPCs, describing the curves' characteristic features, and computed the corresponding FPC scores. Fifteen FPCs were retained from the procedure, which was sufficient to explain > 99% of the variance in all the 32 data sets generated above (Section 4.2.3). Separately, 50 FPCs were retained to evaluate the

impact of retaining no more than 15 FPCs. When registration was used above, FPCA was run on the time-warp curves for each jump, where three 'temporal' FPCs were retained. The FPCs for the VGRF curves were called 'amplitude' FPCs for clarity. No varimax rotation was performed, so issues with multicollinearity could be avoided in the subsequent linear models.

4.2.5 Regression models

Linear regression models were developed to estimate peak power and jump height, respectively. Using the FPC scores as predictors, the linear models took the general form:

$$y_i = b_0 + \sum_{j=1}^{J} b_j x_{ij} \tag{4.1}$$

where y_j is the performance variable (peak power or jump height) for the i^{th} jump, b_j are the model parameters to be determined and x_{ij} are the FPC scores above. The SAS GLMSELECT procedure chose which components to retain in the model using stepwise 'competitive' selection. At each step, the procedure added the component that produced the largest improvement in the Schwarz-Bayesian information criterion (SBC). The competitive element meant that the procedure could remove a previously selected component if, in a later step, its continued inclusion was detrimental to the SBC statistic. Once the model was finalised, the standardised regression coefficients reflected its relative importance of each FPC.

4.2.6 Classification model

A logistic regression model classified the jumps into those performed with and without arm swing. This model tested how well FPCA could identify features that discriminated between the two jump types. This model was intended as a further test of the FPC's ability to represent key movement characteristics that relate to an intentional change in movement strategy, which in this case involves arm swing. Although not directly related to the aim of predicting the performance outcome using a regression model, it was considered worthwhile to include a classification model

since those types of model are more widespread in the machine learning literature. Accordingly, the logistic model took the general form:

$$\text{logit}(\pi_i) = a_0 + \sum_{j=1}^{J} a_j x_{ij} \tag{4.2}$$

where π_j is the probability the jump incorporated arm swing, based on the binomial distribution. As above, x_{ij} are the FPC scores, while a_j are the model coefficients to be determined using the statistical procedure. The SAS LOGISTIC procedure chose the components using a similar stepwise selection based on Fisher scoring. The competitive-type selection was based on the Wald chi-square test statistic. Components were selected when $p < 0.001$ and were retained, provided the p-value remained below this threshold. The model's classification accuracy was given by the concordant rate, equivalent to the area-under-curve (AUC) of the ROC curve (Receiver-Operator Characteristic).

4.2.7 Cross-validation procedures

Monte Carlo Cross Validation (MCCV) provided robust estimates for the model parameters and fit statistics based on 1000 iterations. The data was divided randomly into training and validation sets of equal size (two-fold design) for model selection (Section 2.6.1). The data was split along participant lines rather than by individual jumps ensuring jumps from the same participant could not be in both the training and validation sets. The SBC statistic used above in selecting the regression models was evaluated on the validation set (Section 4.2.5). The standardised coefficients were averaged over all repetitions, allowing 90% confidence intervals to be determined without making assumptions about the distribution. A component's frequency of occurrence in the final model indicated whether it was a generalisable characteristic.

4.2.8 Computer processing

The functional data analysis, including functional smoothing, registration, FPCA and VGRF curve re-construction, was implemented with custom-written code in Matlab R2020a, which called an FDA code library (Ramsay, 2017). The statistical procedures were carried out using SAS 3.8, which used the Matlab-processed data.

4.3 Results

4.3.1 Time normalisation and registration techniques

The smoothed VGRF curves had an RMSE of 7.60×10^{-4} BW with respect to the raw VGRF data when averaged across every point of every jump. The cross validation error was 3.75×10^{-4} BW, which is negligible when considering the calibration error was 5.99×10^{-4} BW (Appendix D). The jump height RMSE, the difference when calculated using smoothed data compared to raw data, was 2.96×10^{-5} m, which for all practical purposes is negligible.

The models based on the PAD data sets were generally able to explain a higher proportion of the variance in peak power and jump height than the LTN models (Figure 4-2). The LTN models tended to benefit more from curve registration overall, but this was not sufficient to produce a better fit than the PAD models. The registered PAD models gained more from the contribution made by the temporal FPCs, implying a corresponding drop in the contribution from the amplitude FPCs as the PAD models did not tend to benefit from registration.

The best model for predicting peak power was based on the unregistered PAD data set (P1), explaining 98.4% of its variance. Table 4-1 presents the results for the best models. The best model for jump height (P2) benefited from curve registration, raising r^2 from 91.4% in the unregistered case to 94.6%. The LTN models explained ≤ 89.4% of the variance in peak power and ≤ 74.8% of the jump height variance. The best classification model correctly distinguished between jumps with and without arm swing 87.0% of the time. This model was the same unregistered PAD model (P1) that produced the best peak power model.

The landmark for the start of the propulsion phase (L_3) was the best individual landmark across all PAD models, featuring prominently in the top four models (Table 4-2). For the peak power models, the peak power landmark (L_4) also appeared three times in the top four. A similar picture emerged for the LTN models (not shown).

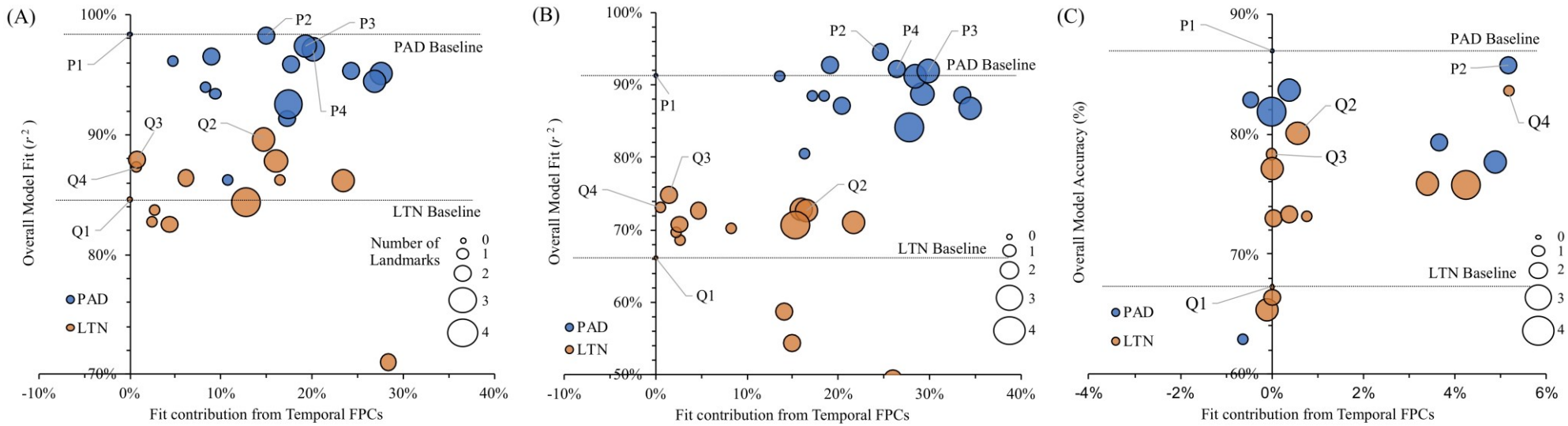


Figure 4-2. Fit of the models for the PAD and LTN data sets with different combinations of landmarks used in curve registration: (A) Peak power model; (B) Jump height model; (C) Classification model. The horizontal axis reveals the contribution made by the temporal FPCs towards the model's overall fit on the vertical axis. The size of the bubble indicates how many landmarks were used in registration.

Table 4-1. Model fit for the selected models identified in Figure 4-2 for peak power, jump height and jump classification.

Model	Data set	Landmark Registration				Peak Power	Jump Height	Classification
		L_1	L_2	L_3	L_4	r^2	r^2	Accuracy
P1	PAD	×	×	×	×	**98.4%**	91.4%	**87.0%**
P2	PAD	×	×	✓	✓	98.2%	**94.6%**	85.8%
P3	PAD	✓	×	✓	✓	97.4%	92.0%	83.7%
P4	PAD	×	✓	✓	✓	97.1%	92.0%	77.6%
Q1	LTN	×	×	×	×	84.6%	66.2%	67.3%
Q2	LTN	×	✓	✓	✓	**89.6%**	72.6%	80.1%
Q3	LTN	✓	×	✓	×	87.9%	**74.8%**	78.3%
Q4	LTN	×	×	✓	×	87.3%	73.1%	**83.6%**

L_1: VGRF minimum; L_2: Power minimum; L_3: Start of propulsion phase; L_4: Power maximum
The best models for PAD and LTN data sets are highlighted in bold.

Table 4-2. Individual landmark contributions to model fit/accuracy averaged over all the registered PAD models. The number of times each landmark appears in the top 4 models is also shown in case averaging across poor models biases the results.

		L_1	L_2	L_3	L_4
Peak Power	Mean Fit (r^2)	84.2%	94.5%	**96.1%**	84.3%
	Number of Top 4 Rankings	1	1	**3**	**3**
Jump Height	Mean Fit (r^2)	79.9%	88.5%	**90.9%**	79.7%
	Number of Top 4 Rankings	1	2	**4**	2
Classification	Mean Accuracy	65.5%	76.2%	**80.5%**	69.0%
	Number of Top 4 Rankings	1	1	2	**3**

L_1: VGRF minimum; L_2: Power minimum; L_3: Start of propulsion phase; L_4: Power maximum
The best fit / accuracy are highlighted in bold, along with the strongest presence in the top four models.

In conclusion, the models based on the PAD method without curve registration (P1) explained the highest proportion of the peak power variance and the second highest in terms of jump height. Using the same data set, the classification model achieved the highest accuracy. Therefore, the P1 data set was selected for further analysis.

4.3.2 VGRF Components

The first eight FPCs that make up these VGRF curves are presented in Figure 4-3, revealing each component's characteristic variation. The deviation from the mean curve shows the direction of increasing jump performance, or in the case of the classification model, the effect of arm swing. Jump performance includes peak power and jump height since they had the same signs as the FPC coefficients except for FPC1, which was usually excluded from the peak power model (section 4.3.3 below).

FPC1 described a higher peak in VGRF at the start of a more prolonged propulsion phase associated with a greater jump height, but it did not contribute to peak power (Figure 4-3A). There was also a mixture of amplitude and phase variation that was only observed in this FPC. In the classification model, FPC1 was excluded, which explains the absence of a dotted line in this plot.

FPC2 also described variations in the amplitude range associated with peak power and jump height (Figure 4-3B). In contrast, arm swing worked to diminish the amplitude range in VGRF. FPC3 described a final peak in the VGRF toward the end of the propulsion phase before take-off (Figure 4-3C) that was associated with both jump performance measures. For FPC4, the timings were similar to FPC2, but the directions reversed between the first and second fluctuations (Figure 4-3D). The next four components, FPC5 to FPC8, make various subtle changes to the VGRF curve that had a modest influence on jump performance.

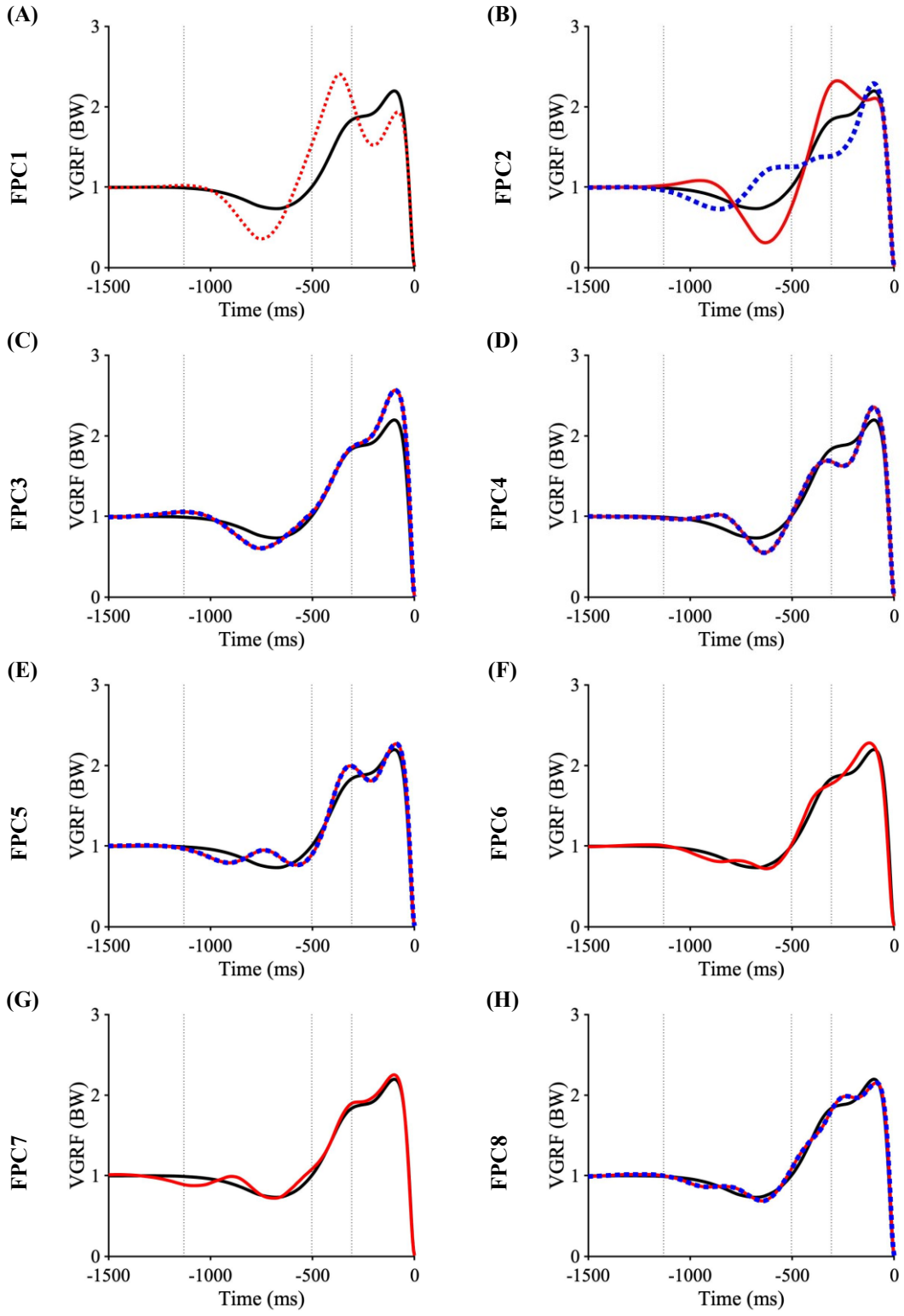


Figure 4-3. Functional Principal Components, FPC1-FPC8, with deviations from the mean curves (solid black line). Solid red line = deviation when peak power in higher than the mean. Dashed blue line = deviation indicative of a jump performed with arm swing. Both lines overlap in several plots, indicating arm swing also contributes to peak power. Deviation 2 × SD. Vertical dotted lines indicate the phase boundaries between (from left to right) the unweighting phase, the braking phase and the propulsion phase.

4.3.3 Model fit

The central estimates from cross-validation were almost identical to the original estimates from the models above (section 4.3.1). However, the lower confidence limit for peak power (10^{th} percentile of models) dropped to 92.9% (Table 4-3). The confidence intervals were narrower for the jump height models, showing those models were less prone to overfitting. When making predictions based on the validation set, the RMSE central estimates were 1.40 $W \cdot kg^{-1}$ and 0.033 m for peak power and jump height, respectively. These errors were slightly higher than the RMSE based on the training set (0.99 $W \cdot kg^{-1}$ and 0.027 m, respectively).

Table 4-3. Model fits for the unregistered PAD models averaged over 1000 repetitions of cross-validation. The central estimates are shown with 90% confidence intervals in brackets.

Model	† Explained Variance / ‡ Accuracy	Validation RMSE
Peak Power †	98.3 [92.9, 98.6] %	1.40 [1.19, 1.61] $W \cdot kg^{-1}$
Jump Height †	92.0 [90.3, 93.6] %	0.033 [0.030, 0.037] m
Classification ‡	88.8 [85.6, 91.9] %	n/a

† Explained variance is r^2.
‡ Classification accuracy is the concordant rate.

When the number of retained FPCs was increased from 15 to 50, the average validation RMSE for the peak power models only dropped slightly to 1.19 $W \cdot kg^{-1}$ (Table 4-4). In contrast, the average error for the jump height models was cut from 0.033 m to 0.015 m, whilst the classification accuracy rose from 88.8% to 95.0%. Using 50 FPCs minimised the approximation error of retaining 15 FPCs, providing an indication of how good the linear models could be.

Table 4-4. Model fits based on retaining 50 FPCs for the unregistered PAD models, averaged over 1000 repetitions of cross-validation. The central estimates are shown with 90% confidence intervals in brackets.

Best Models	† Explained Variance / ‡ Accuracy	Validation RMSE
Peak Power †	99.3 [99.1, 99.5] %	1.19 [1.01, 1.39] $W \cdot kg^{-1}$
Jump Height †	99.2 [98.9, 99.5] %	0.015 [0.013, 0.016] m
Classification ‡	95.0 [92.3, 97.6] %	n/a

† Explained variance is r^2

‡ Classification is the concordant rate

4.3.4 Model components

The stepwise selection process produced a range of models over the 1000 cross-validation repetitions that comprised different sets of FPCs. Most of the FPCs appeared in the peak power models in every repetition (FPC2–FPC6, FPC8, FPC10 and FPC12: Figure 4-4A), whereas fewer and some different FPCs were always present in the jump height models (FPC1–FPC3, FPC5, FPC6, FPC8 and FPC9: Figure 4-4B). The classification models were sparse in comparison, with only FPC4 always being present (Figure 4-4C), while FPC2, FPC3 and FPC5 appeared almost every time (in ≥ 950 occasions). These are the FPCs that were strongly associated with arm swing. In contrast, for the performance models, multiple FPCs were linked to peak power or jump height.

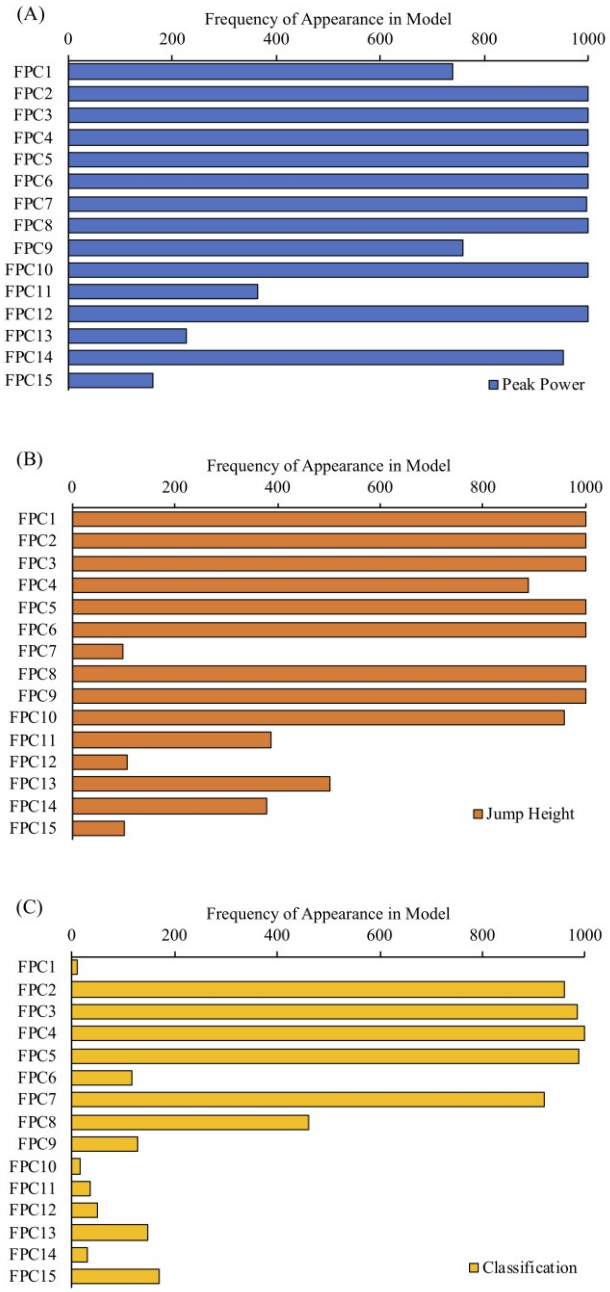


Figure 4-4. Frequency that FPCs appear in the unregistered PAD models based on 1000 iterations of 2-fold cross-validation. (A) Peak power; (B) Jump height; (C) Classification (arm swing).

The proportion of the variance in the jump performance outcome variables showed that FPC3 accounts for by far the largest proportion of the variance in peak power and jump height (central estimates of 72.9% and 57.7%, respectively – Figure 4-5). The next best FPCs explained < 10% of the outcome variances (FPC8 and FPC5). In contrast, the curves'

variance accounted for by FPC3 was only 8.6%, whilst FPC1, which always accounts for the largest curve variance, explained roughly half the variance (50.7%).

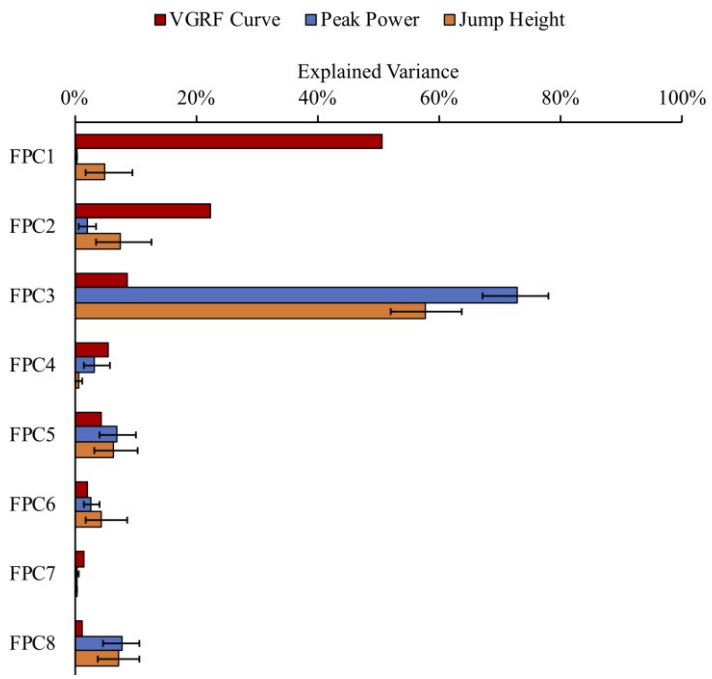


Figure 4-5. Variance in the VGRF curves explained by the unrotated FPCs for the unregistered PAD data set compared to the variances in the two jump performance measures. Error bars indicate 90% confidence intervals estimated from cross-validation. FPC9–FPC15 are not shown for clarity as they made minimal contributions.

The magnitudes of the standardised parameter estimates indicate the FPC's contribution, irrespective of the different scales involved (Figure 4-6). FPC3 had the largest parameter estimate of the performance models, with the positive sign indicating a higher FPC3 improves jump performance. FPC4 had the largest magnitude in the classification model, and being negative indicated that arm swing tended to *decrease* the FPC4 score.

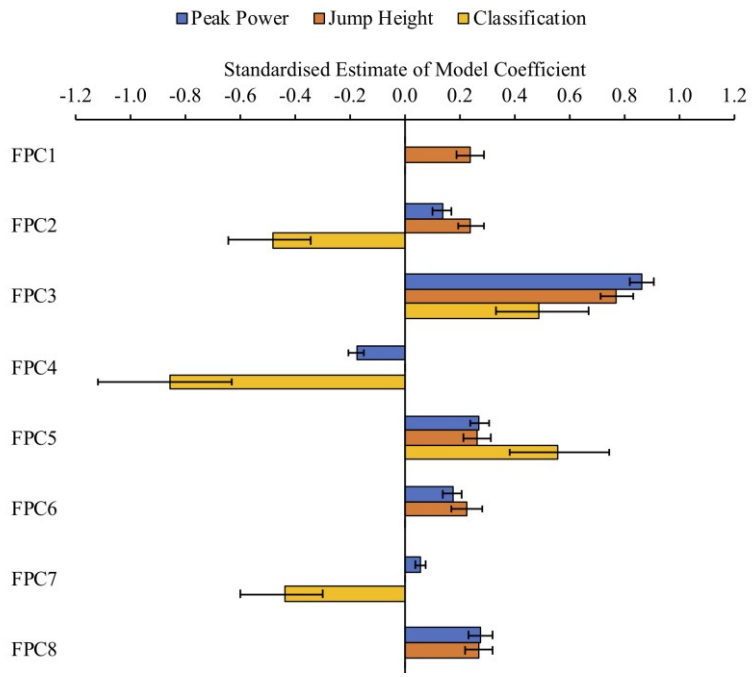


Figure 4-6. Standardised estimates for the FPC coefficients in the models based on the unregistered PAD data set. For clarity, the bars shown are only for FPCs appearing in models > 900 times out of 1000 repetitions of cross-validation. Error bars indicate 90% confidence intervals.

No such comparisons can be made for the classification models because predictive accuracy is not additive. Instead, the standardised odds ratios revealed the relative importance of each FPC to arm swing (Table 4-5). A jump with an FPC4 score one standard deviation *lower* than that of another jump would indicate that the jump in question was 4.5 times more likely to have involved arm swing in comparison. (The standardised odds ratio was 0.22 for a one SD increase.) The likelihood worked in the opposite direction for FPC3 and FPC5 such that a score one SD *higher* would make it more likely that the jump incorporated arm swing. Overall, the variance in the FPC4 was the biggest indicator for the presence or otherwise of arm swing.

Table 4-5. Standardised odds ratios (OR) for selected FPCs in the classification model, which indicate how many times more likely the jump involved arm swing for every one-SD increase in the FPC score. Odds ratios < 1 indicates such an increase makes it more likely that the jump did *not* involve arm swing. In those cases, the inverted standardised OR is shown to allow comparison of the relative magnitudes. Central estimates are shown with 90% confidence intervals in brackets based on cross-validation.

	Standardised OR		Inverted Standardised OR	
FPC2	0.42	[0.30, 0.54]	**2.37**	[1.86, 3.31]
FPC3	**2.66**	[1.80, 3.76]		
FPC4	0.22	[0.12, 0.32]	**4.50**	[3.12, 8.06]
FPC5	**2.90**	[2.00, 4.09]		
FPC7	0.45	[0.34, 0.57]	**2.20**	[1.75, 2.92]

FPCs selected appeared > 900 times in the model over 1000 repetitions of cross validation.
The standardised (or inverted standardised) odds ratios > 1 are highlighted in bold to make it easier to compare their magnitudes.

4.4 Discussion

The aim of this chapter was to investigate the efficacy of FPCA-type models in vertical jumping based on gold-standard VGRF data before applying this approach to accelerometer data. Two different time normalisation techniques were considered to standardise the time series' length by either resampling using interpolation (LTN) or by padding out the time series to the required length (PAD). In addition, curve registration was evaluated to determine if it could benefit the models. These methods were evaluated by considering: (1) the goodness of fit of the resulting models (explained variance or predictive accuracy); (2) how robust the models' predictions were when applied to validation data over many sub-samples; and (3) whether the characteristic features of the VGRF curves related to known neuromuscular factors. Although the prediction of peak power is of primary interest, models of jump height were developed as a useful comparator when interpreting the results. The findings from this investigation will help to determine the best methods to use on the accelerometer data.

4.4.1 Goodness of fit and predictive error

The most obvious way to judge the quality of the model, and arguably the best, is to assess its predictive accuracy. If the model correctly captures the intrinsic qualities that govern the outcome, it should predict it with high fidelity. On this measure, the simple linear models based on the FPCs were able to explain a very high proportion of the variance in peak power (98.3%) and jump height (92.0%), which translated to validation RMSEs of 1.40 $W \cdot kg^{-1}$ and 3.3 cm, respectively. The logistic regression model classified roughly 7 out of 8 jumps correctly (88.8%). The jump height model was more accurate than a model based on discrete 'features', which had an RMSE of 4.8 cm (Dowling & Vamos, 1993), as Richter et al. (2014a) had previously found for FPC-based models.

Richter et al. (2014a) and Moudy et al. (2018) reported fits of 98.8% and 86.1%, respectively, for their ACP models. Those studies used the flight-time definition, making jump height easier to predict because the linear model becomes a sum of net impulses. In contrast, the current study used the work-done definition of jump height, which is more representative of active play in many sports such as volleyball and basketball. However, this jump height definition has a more complicated relationship with the impulse as it includes the so-called 'stretch height' at take-off, plus the height gained in flight.

For a direct comparison with Moudy et al. (2018) and Richter et al. (2014a, 2014b), it is necessary to re-define jump height as the height gained after take-off, as those authors have done. The revised model explains 98.9% [98.7%, 99.0%] of the variance in this alternative measure, which is almost the same as Richter's model. Richter et al. (2014b) and Moudy et al. (2018) quote predictive errors a mean average errors (MAE) of 0.59 cm and 1.37 cm, respectively. As they did not employ cross-validation, the more appropriate comparison is with the revised training RMSE, which for the current study was 0.9 cm [0.8 cm, 1.0 cm]. For reference, RMSE will always be equal to or larger than MAE. These comparisons demonstrate that it is possible to achieve a comparable level of accuracy in predicting jump height without employing ACP.

However, it is the models' predictions of peak power that is of interest in this thesis. The validation RMSE in the current study was 1.40 $W \cdot kg^{-1}$ (based on 15 FPCs). These predictive errors are below the typical intra-day CV of 1.75 $W \cdot kg^{-1}$, averaged across four studies (Section 2.2.2), although not substantially so. Unlike the jump height and the

classification models, the accuracy of peak power predictions did not improve markedly when 50 FPCs used in the model. The predictive error for this extended model was slightly lower at 1.19 $W \cdot kg^{-1}$. The differences with the training RMSE (0.99 $W \cdot kg^{-1}$ and 0.70 $W \cdot kg^{-1}$, for 15 and 50 FPCs respectively) were not particularly large, suggesting that overfitting, while present, was not a big factor. Out of the 50 FPCs entered into the model, only 14 FPCs were selected more often than 900 out of 1000 iterations. This finding shows that the model did not prefer FPCs describing minimal VGRF variations.

The above evidence suggests that peak power is harder to predict than jump height, perhaps because it depends on a narrower timeframe when the forces acting on the body vary rapidly. In contrast, jump height depends on the cumulative effect of force. Thus, the peak power model's parameters were more sensitive to small variations in the data. In comparison, Bobbert et al. (1987) showed that peak power was sensitive to deliberate changes in drop jumping technique, while the total work done (proportional to the jump height) remained the same. Therefore, it may be more appropriate to develop a more sophisticated model than the linear model used in the current investigation. Therefore, the next chapter will consider a range of different machine learning models when applying the current FPCA techniques to the accelerometer data.

4.4.2 Time normalisation and curve registration

Time normalisation and curve registration had a pivotal effect on the models because they altered the VGRF curves' shape and hence, the FPCs themselves. Padding out the start of the VGRF curve to reach a standard length always produced a better performance model compared to LTN. Rescaling the time domain through registration undermined the relationship with impulse, reducing the model's predictive accuracy. In the classification case, no direct link with impulse exists, so the differences between the PAD and LTN models were much less apparent. For the unregistered PAD data set (P1), there could still have been misalignments between the VGRF curves, but relatively small amounts of phase variance may be tolerable when there is a need to preserve the time domain (Ryan et al., 2006; Donoghue et al., 2008; Epifanio et al., 2008; Crane et al., 2010). Registration may have gone some way towards 'correcting' the distortions from time normalisation, given

that the LTN-based performance models benefited much more from registration than the PAD models (7 cases for peak power models and 11 for jump height models).

The most efficacious landmark was located at the start of the propulsion phase (L_4), as it appeared so often in the best models. Moudy et al. (2018) found that the same landmark produced the best jump height models. Other studies have reported that the propulsion phase duration is the least invariant (Oddsson, 1987; van Ingen Schenau, 1989), so it might be expected that the curves would need the smallest adjustment to align about this landmark. However, the plots in Figure 4-3 shows that the VGRF undergoes the highest rate of change during the latter part of the braking phase, making the model's parameters more sensitive to small changes. Landmark registration helped improve the models' fit by decomposing curve variance into its amplitude and temporal components. Using continuous time-warping functions is more effective than traditional dynamic time-warping which depend on discrete measures (Marron et al., 2015). The current analysis shows that some benefit can be gained from registration in some circumstances. It should be considered when applying the FPCA technique to the accelerometer data (Chapter 6) once a satisfactory accelerometer model has been identified (Chapter 5).

4.4.3 Cross validation

Cross validation played an essential role in assessing the models as it provided robust estimates of predictive error. The FPCs appearing in the model every time were those predictors that were more likely to generalise well to unseen data, lessening the likelihood of overfitting. Conversely, the FPCs that explained only a very small proportion of the overall variance, describing nuanced characteristics, appeared infrequently illustrating how fitting to extraneous features does not generalise well to other data sets. These observations confirm the value of using cross validation to identify the predictors that generalise well across data sets. It should also be noted that the models' predictors were based on FPCA, which treated all curves equally. The cross validation procedure in SAS was performed after FPCA was run in Matlab. Hence, the FPCs themselves were defined from training and validation data sets combined. This limitation will be addressed in the next chapter with a Matlab-only implementation.

4.4.4 Interpretation of the FPCs

The usefulness of a model can extend beyond the accuracy of its predictions if it can provide insight to the investigator on how the predictions were made. The models based on FPCA can offer insights on the biomechanics of vertical jumping on two levels: (1) the relative importance of each FPC can be examined; and (2) the shapes described by the FPCs are intuitive, allowing meaningful interpretation. In terms of performance, FPC3 was the single most important feature, describing an initial variation in unweighting phase along with a peak in the propulsion phase, just before take-off. The FPC3 score explained most of the variance in peak power and jump height (72.9% and 57.7%, respectively), dwarfing the contributions made by the other FPCs (Figure 4-5). Oddsson (1989) also found that the peak VGRF immediately before take-off was the single best predictor in his jump height model for CMJ_A, explaining 44% of the variance. Notably, FPC3 from the current study accounted for a larger proportion of the variance than the absolute VGRF peak value. These results support the principle of late force production first proposed by Hochmuth & Marhold (1977). As discussed in Section 2.2.5, the highest final velocity is achieved (take-off velocity in jumping) when maximum force is applied toward the end of an explosive movement rather than at the beginning. A recent study reported similar findings, revealing the difference between low- and high-performing jumpers lay in the late propulsion phase, between 79% and 97% of the time-normalised VGRF curve (R. Kennedy & Drake, 2018). Moudy et al. (2018) also reported that the key phase for jump height was 84-100% of the time-normalised VGRF curve.

It naturally follows that producing high VGRF at higher velocities requires considerable power. That is why FPC3 was also strongly associated with peak power, as well as with jump height. Moreover, the FPC3 peak was almost coincident with power production reaching its peak. It is reasonable to infer that neuromuscular factors may underlie the differences in the FPC3 score. FPC3 described the variation in the VGRF over roughly the final 280 ms before take-off, which is consistent with the period when the triceps surae contribute strongly to the jump (Bobbert et al., 1986). FPC3 may also be a consequence of the biarticular muscles' action transferring power from the hip extensors to the ankle plantar flexors over this period (Jacobs et al., 1996; van Ingen Schenau, 1989).

Researchers have also observed how some force-time curves in jumping have one peak while others have two, often in connection with arm swing (Payne et al., 1968; Shetty & Etnyre, 1989; Miller & East, 1976). Interest was renewed recently in these patterns with studies reporting no clear advantage in terms of jump height between unimodal and bimodal curves (R. Kennedy & Drake, 2018), along with inconsistency in the VGRF curves' modality in jumps from the same individual (Lake & McMahon, 2018). Cockcroft et al. (2019) proposed a method for the objective classification of modal curves based on a detailed analysis of discrete measures of maxima and minimum force levels over the propulsion phase. The current study offers an alternative framework based on continuous features described by the FPCs. It cannot classify curves as Cockcroft et al. (2019) does, but it offers a fresh perspective into the underlying movement characteristics behind curve modality.

Finally, it is worth noting that FPC1 and FPC2 explained 73.0% of the VGRF variance but only accounted for 12.1% of the jump height variance and had almost no influence on peak power, accounting for just 1.5% of its variance. These observations illustrate that although there may be considerable variation in the VGRF curve, such variations may not impact the jump performance. Such variation may be attributed in part to the degrees of freedom available to the jumper in how they execute the jump (Dowling & Vamos, 1993; J. Jensen et al., 1989) and may reflect natural movement variability (Glazier et al., 2006; Button et al., 2003; Davids et al., 2003; Kelso, 1995). Such variability presents a challenge in estimating peak power and jump height from accelerometer signals. The results show that only a small proportion of the variance in the VGRF is relevant to performance, and it may be reasonable to suppose that the same will apply to the accelerometers signals. If so, the challenge will be to identify those features in the accelerometer data that are relevant to the generation of peak power.

From the above discussion, it should be clear that FPCA allows meaningful qualitative interpretation of the FPCs in an intuitive way. Whilst other authors have used varimax rotation to aid interpretation, varimax was not necessary here, nor was it desirable as such rotations would introduce correlations between the previously independent FPC scores. Whilst linear models are generally tolerant of moderate correlations between the predictors, there were several instances of multicollinearity associated with the varimax FPCs when

they were tested. The multicollinearity reached such high levels that it would have cast doubt on the models' robustness ($r > 0.9$ and VIF > 10).

4.4.5 Conclusions

FPCA has proven itself to be an effective means for extracting characteristic features from the VGRF data. The robust models could predict performance outcomes and classify jump types with a high degree of accuracy when tested on unseen validation data. The regression models demonstrated that FPCA could pick out characteristics related to peak power and jump height. The two measures served to test this capability in different ways. Although classification was not the aim of this thesis, the logistic model also showed that the FPCs could be used to distinguish between two different movement patterns. Hence, the FPCs describe biomechanically meaningful aspects of the jumping movement. These results indicate FPCA has the potential to be a useful technique for accelerometer data. As an unsupervised learning technique, FPCA has no *a priori* assumptions of what the key features should be, nor is it guided by the intended performance metric. Nevertheless, the FPCA-based models achieved high levels of accuracy, perhaps because these features explained many aspects of jumping so elegantly.

This chapter also established that padding out the time series yields more accurate models than when using time normalisation. Padding at the start ensured the curves remained in reasonably close alignment near take-off, the critical period when peak power is achieved. Curve registration could not improve alignment much further. Registration marginally improved the jump height model, but otherwise, it was detrimental to the PAD models. It did assist the LTN models only because time normalisation had left the curves more misaligned with one another. Therefore, padding is preferable to time normalisation, while curve registration may benefit if padding is not sufficient to properly align the curves. Thus, with VGRF gold standard data, FPCA has proven itself to be an effective feature extraction method and can therefore be applied with confidence to accelerometer data. The remainder of this thesis will now seek to develop and investigate FPCA-based models based on accelerometer data.

CHAPTER 5. ACCELEROMETER MODEL SELECTION OPTIMISATION

5.1 Introduction

This thesis so far has focused on the VGRF data, using it to compute the criterion jump performances (Chapter 3) and evaluate FPCA as a feature extraction method (Chapter 4). As the previous chapter established, the FPCs can represent key characteristics related to the performance outcome. This approach can now be applied to answer the second research question, *which machine learning algorithms are better suited to predicting peak power in vertical jumping based on the FPCA characteristics of body-worn accelerometer data?* The accelerometer data will come from one of four sensors attached to either the lower back (LB), upper back (UB) or the shanks (LS/RS) (Section 3.2.4). One of those attachment sites may be better suited than the others for predicting peak power in vertical jumping. Similarly, higher accuracy may be achieved for one jump type over the other (CMJ_{NA} vs. CMJ_{A}). Hence, there are eight different data sets for each combination of anatomical location and jump type. Determining at this initial stage which data set is more likely to yield accurate predictions will streamline the investigation in subsequent chapters.

Models based on accelerometer data may need to be more sophisticated than the linear model for VGRF data, given the inherent difficulties in using body-worn sensor data to predict athletic performance (Sections 2.3.4–2.3.5). Hence, the models considered in this chapter should encompass a range of possible algorithms, including parametric and non-parametric models that take different approaches to controlling overfitting. They should be chosen based on their generalised predictive accuracy so the models can be applied with a reasonable level of confidence to new data that is gathered in the same way. Nested cross validation provides the framework to achieve this as it yields unbiased estimates of the model's predictive error (Section 2.6.2). Operating inside the NCV inner loop, optimisation will play a central role in model selection (Section 2.6.3). As part of this process, it will help understand the key factors that influence model accuracy rather than treating it as a black box. Throughout this investigation, computation cost will be a consideration as many

thousands of subsamples will be needed to produce an unbiased estimate of the model's generalised error.

The aims of this chapter are therefore to:

- Determine which combination of jump type and sensor location provides the data set best suited to making accurate peak power predictions;
- Select and optimise suitable machine learning algorithms using the nested cross validation framework; and
- Obtain generalised estimates of the selected model(s) predictive error from this first stage of accelerometer model development.

5.2 Methods

5.2.1 Overview

The methods in this chapter will be applied throughout the rest of this thesis concerning accelerometer data. They begin by describing the data preprocessing needed to convert the accelerometer signals into smooth continuous functions and then extract the FPC-type features. A shortlist of promising models is drawn up based on K-fold cross validation before nested cross validation (NCV) is applied. A novel optimisation procedure is introduced based on a Bayesian surrogate model trained on observations from a random search, which is progressively constrained towards the optimal region. Particle Swarm Optimisation (PSO) finds the surrogate model's global optimum, which corresponds directly to the optimal parameter values for the algorithm in question (J. Kennedy & Eberhart, 1995). A final model is determined for each of the shortlisted algorithms by aggregating the multiple models produced in the NCV procedure using the bagging technique (Breiman, 1996). The final models are then evaluated to determine the generalised predictive error using the Monte Carlo Cross Validation (MCCV). The final models are retrained on the full data set and then tested on the holdout set as a further independent check.

5.2.2 Data preprocessing

The calibrated 3D accelerometers signals obtained from Chapter 3 comprised $N = 275$ jumps without arm swing from 60 participants. The accelerometer signals, $\mathbf{a}_n(t)$, were converted into resultant time series (L_2-norm), denoted by $\mathcal{D}_0 = \{|\mathbf{a}_n(t)|\}$, where the index $n \in \{1, \ldots, N\}$ identifies each jump. Using the resultant signal simplified the analysis at this initial stage by considering a single waveform for each jump, although in doing so, it removed information related to the sensor's orientation. The accelerometer time series were time-shifted backwards 100 ms (25 time intervals) to synchronise them with the VGRF data, based on the study into the accelerometers' temporal validation (Appendix B). The time series were then truncated to fit a prescribed time window covering the 2 s immediately before take-off [-2 s, 0 s]. Thus, each time series comprised 501 points, given the 250 Hz sampling frequency.

The conversion into smooth continuous functions used the same method as before (Section 4.2.2), employing 100 b-spline basis functions to obtain the same density of overlapping functions as the VGRF curve. Fourth-order basis functions were sufficient to avoid discontinuities in the second derivatives of the smoothed signals (i.e. a smooth curvature), one order lower than the VGRF basis functions. Accordingly, the roughness penalty was applied one level lower at the second order. The roughness penalty, $\lambda = 10^2$, was determined by Generalised Cross-Validation (GCV) (Ramsay & Silverman, 2005). With this level of smoothing, 100 b-splines was close to the optimal amount for efficiently retaining information from the accelerometer data: 94.8 degrees of freedom out of a possible 100 (Ramsay & Silverman, 2005). Whilst more b-splines would have provided greater flexibility, the computational costs would have been much higher. Fifteen FPCs were retained from FPCA without varimax rotation, following the same approach as with the VGRF data, as they accounted for > 99% of the curve variance. Chapter 7 will determine the optimal number of retained components that minimises predictive error.

The FPCs were defined based on the training set alone, unlike in the previous chapter, to ensure full separation between the training and validation sets in respect of the definition of the features themselves. The Matlab FDA code library does not support this procedure, so a custom function was developed to calculate the validation FPC scores using the training

FPCs. The calculation involved projecting each validation acceleration curve onto the training FPC using the inner product.

5.2.3 Candidate accelerometer models

The FPC scores served as the predictors in the models. The model algorithms covered parametric and non-parametric methods, neural networks and a tree ensemble (Table 5-1). Since feature definition was based on FPCA, more sophisticated deep neural networks with their convolutional layers designed to learn features in the data were not applicable, nor were Long/Short-Term Memory (LSTM) networks that take the raw time series as their input (Hochreiter & Schmidhuber, 1997). No feature selection methods were employed, as mentioned above, such as stepwise selection from the previous chapter. It was preferable to focus on the algorithms' efficacy rather than introducing additional complexity at this stage.

Many of these algorithms were available in Matlab's Regression Learner App (R2020a), which offers a convenient way to evaluate several different types of models. However, the app did not support a form of grouped cross-validation that allowed the data to be split along participant lines, as was done in the previous chapter (Section 4.2.7). In addition, the app did not allow any form of optimisation, the absence of which may have disadvantaged some algorithms at this first stage of model selection. Instead, the values for the models' hyperparameters were determined using Matlab's hyperparameter optimisation (HPO) feature, the results of which are also presented in Table 5-1. The 'auto' option for HPO was chosen as it was designed to optimise the hyperparameters generally considered to be influential on model performance. The few hyperparameters omitted, including the GPR basis function and standardisation options, reverted to their Matlab default settings. No cross-validation was incorporated into the HPO for the same reason as above. It was preferable to adopt a lightweight approach at this stage, given the large number of models involved. The subsequent optimisation requiring considerable computational resources was reserved for a few shortlisted models.

Table 5-1. Machine learning models for evaluated in this chapter.

Algorithm †	Identifier	Key Hyperparameter Values ‡
Linear Regression: **fitrlinear** (Hastie et al., 2009; Hoerl & Kennard, 1970; Marquardt & Snee, 1975; R. Tibshirani, 1996)	LR	Lambda (λ_{LR}) = 0 Solver = Least Squares
	LR-RDG	Lambda (λ_{LR}) = 0.1 Solver = Least Squares
	LR-LSS	Lambda (λ_{LR}) = 0.01 Solver = SVM
Support Vector Machine: **fitrsvm** (Boser et al., 1992; Drucker et al., 1996; Smola & Schölkopf, 2004)	SVM-L	Kernel = Linear Box Constraint (BC) = 0.001 Kernel Scale (KS) = 0.4 Epsilon (ϵ) = 2.75
	SVM-G	Kernel = Gaussian Box Constraint (BC) = 220 Kernel Scale (KS) = 270 Epsilon (ϵ) = 2.65
Gaussian Process Regression: **fitrgp** (Bishop, 2006; Rasmussen, 1999; Rasmussen & Williams, 2006)	GPR-SE	Kernel = Squared Exponential Sigma (σ) = 2.1
	GPR-M52	Kernel = Matérn 5/2 Sigma (σ) = 2.3
Feedforward Neural Network: **feedforwardnet** (Hastie et al., 2009; Sapna, 2012)	NN-5	Hidden Nodes = 5 Training = Levenberg-Marquardt
	NN-10	Hidden Nodes = 10 Training = Levenberg-Marquardt
Tree Ensemble: **fitrensemble** (Dietterich, 2000a, 2000b)	TR-ENS	Method = Least Squares Boost Learning Cycles = 10 Learning Rate = 0.4 Minimum Leaf Size = 1

† Associated Matlab command in bold.
‡ Determined by Matlab Hyperparameter Optimisation with 'auto' method.

5.2.4 Shortlisting Models

The process to shortlist the accelerometer models (Figure 5-1) involved using a series of grid searches for every combination of algorithm, sensor and jump type (Table 5-2). The validation errors were subject to statistical analysis to identify the best data sets. The models were evaluated based on an estimate of their predictive error using two-fold MCCV with 100 iterations (Molinaro et al., 2005; Xu & Liang, 2001; P. Zhang, 1993). With its equal-sized training and validation sets, two-fold cross-validation minimised the variance in the loss between folds and is recommended for model selection (Breiman & Spector, 1992; Hall & Robinson, 2009; Shao, 1992, 1993; Y. Zhang & Yang, 2015). The standard error in

RMSE with 100 MCCV iterations was found to be 0.05 $W \cdot kg^{-1}$ for this arrangement when averaged across all models (AppendixE.1). As in previous chapters, the partitioning was done by participant rather than by trial. Had this not been done, it would have resulted in an under-estimation of the validation error by up to 0.65 $W \cdot kg^{-1}$ (Appendix E.2). Each subset (training and validation) comprised jumps from 30 participants, ranging from 120 to 152 (137 ± 5) depending on the breakdown on jumps performed by participant (4 or 8) (Section 3.2.2). The small differences in sample size had no discernible effect on the models' predictive errors because minor differences were averaged out over multiple folds.

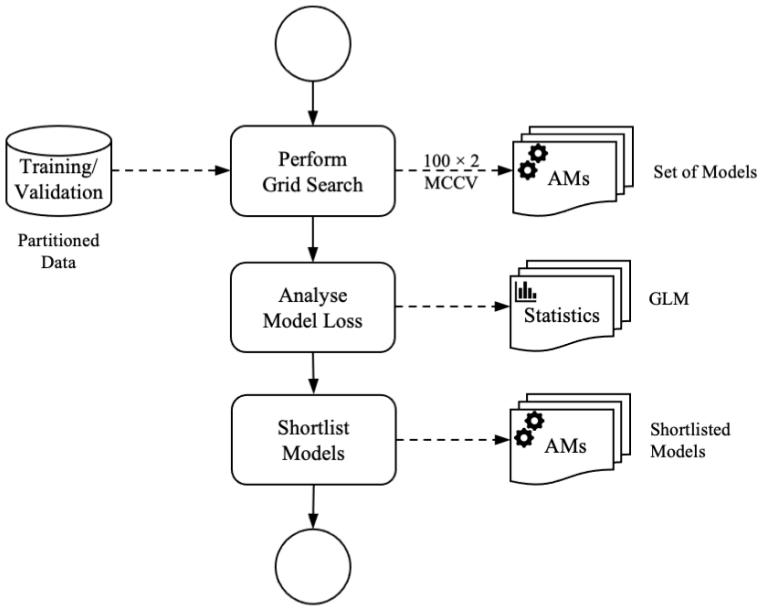


Figure 5-1. Process flow for the first stage of model selection using a grid search and subsequent statistical analysis to shortlist models for further analysis. AM = accelerometer model; GLM = Generalised Linear Model for statistical comparisons of model errors.

Table 5-2. Parameter search ranges for the grid search, which included every combination of sensor attachment site, jump type and algorithm shown below.

Parameter	Search Range
Sensor attachment site †	{LB, UB, LS, RS}
Jump type	{CMJ_{NA}, CMJ_{A}}
Algorithm ‡	{LR, LR-RDG, LR-LSS, SVM-L, SVM-G, GPR-SE, GPR-M52, NN-5, NN-10, TR-ENS}

† LB = Lower Back; UB = Upper Back; LS = Left Shank; RS = Right Shank.
‡ See Table 5-1 for definitions.

Statistical comparisons were made between the models using a generalised linear model (GLM) in preference to ANOVA because the RMSE values across the folds were not normally distributed. The factorial design included model, sensor and jump type with effect sizes determined using partial ω^2, which explained the proportion of the variance not covered by the other variables. Three out of the five algorithm types with lower prediction errors were shortlisted and taken forward to the next stage.

5.2.5 Nested Cross Validation

Cross validation thus far has been used exclusively for model selection. However, in this second stage, cross validation must fulfil the twin objectives of model selection and model evaluation. The purpose of model evaluation is to estimate the model's generalised predictive error to indicate how well the model can be expected to perform on new data sets drawn from the same distribution. Therefore, a nested cross validation design was introduced to separate model evaluation from model selection to guard against biasing the predictive error (Varma & Simon, 2006) (Figure 5-2). The outer cross validation had a 10-fold design, recommended for model evaluation (e.g. Y. Zhang & Yang, 2015), without overlapping validation sets to ensure the validation sets were truly independent (Isaksson et al., 2008). All aspects of data processing, model training and selection were carried out within the outer training partition (Varma & Simon, 2006). The model that emerged was evaluated on the outer validation partition, which had played no part in selecting the model. This procedure was repeated for each outer fold, producing a set of 10 models with their associated independent validation errors. A more generalised model was obtained by aggregating the models using bagging (Breiman, 1996).

In order to accommodate a large number of observations required from the accelerometer model (AM), the number of MCCV iterations was reduced from 100 to 20. This brought a modest increase in the standard error of each CV estimate from 0.05 to 0.12 $W \cdot kg^{-1}$ (AppendixE.1). Emphasis was placed on gathering more observations for greater coverage of the search space than on the precision of individual cross validation estimates. The standard error was effectively reduced during the optimisation when observations were made in the same vicinity. The reduction to 20 iterations of MCCV, a low number of repetitions for a Monte Carlo method, raised the possibility of switching to repeated k-fold CV, i.e. 10 ×2. However, comparisons between the two methods revealed no discernible differences between them in terms of error estimates across all algorithms (Appendix E.3). Therefore, MCCV was retained as the preferred cross validation method. It is a technique that will be used throughout this thesis.

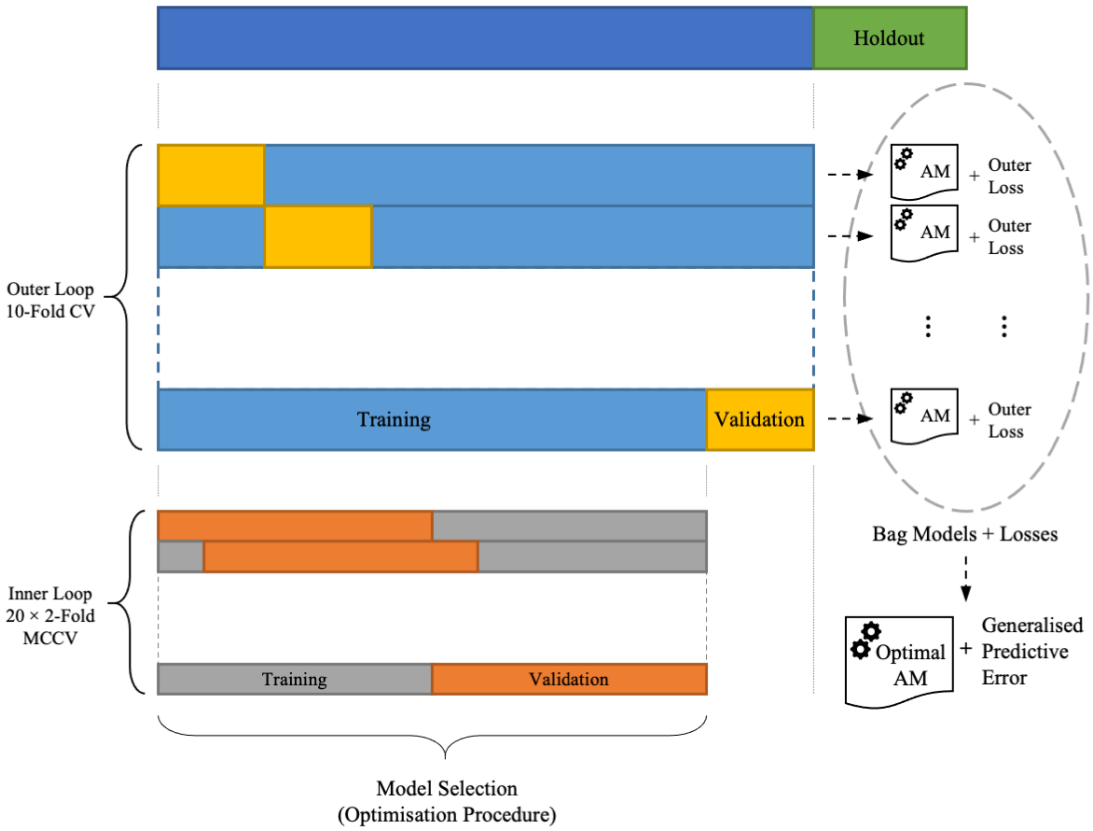


Figure 5-2. Nested cross validation design used in the optimisation procedure in this and subsequent chapters, showing model selection takes place inside the inner loop, producing a model for each outer fold, which are then aggregated using bagging.

5.2.6 Notation

It will be helpful to adopt the following notation to clarify the difference between the data sets in a nested cross validation procedure. Let the *k*-th outer fold be represented by $\mathcal{D}^{(k)}$ where the bracketed superscript identifies the fold, such that $k \in \{1, \dots, K\}$, for which, $K = 10$ in this design. Using subscripts, let $\mathcal{D}_{+}^{(k)}$ and $\mathcal{D}_{-}^{(k)}$ refer to the training and validation sets, respectively. The bracketed superscripts may also identify the inner fold such that $\mathcal{D}_{+}^{(k,l)}$ and $\mathcal{D}_{-}^{(k,l)}$ denote the respective training and validation subsets for the outer training set, $\mathcal{D}_{+}^{(k)}$, where $l \in \{1, \dots, L\}$ identifies the inner fold ($L = 20$).

Thus, FPCA, which was performed on the inner training set (Section 5.2.2), may be defined as:

$$\left[\mathbf{X}_{+}^{(k,l)},\ \Psi^{(k,l)} \right] = \mathrm{FPCA}\left[f(\mathcal{D}_{+}^{(k,l)}) \right] \tag{5.1}$$

where $f(\cdot)$ represents functional smoothing. $\mathbf{X}_{+}^{(k,l)}$ represents the training data used by the algorithm and $\Psi^{(k,l)}$ defines the discretised wavefunctions of the associated FPCs, which applies to $\mathcal{D}^{(k,l)}$ as a whole, including training and validation sets. The validation set was generated by using the inner product to project the smoothed validation curves onto the FPCs (Section 5.2.2), which may be represented symbolically as:

$$\mathbf{X}_{-}^{(k,l)} = \sum \left[f(\mathcal{D}_{-}^{(k,l)}).\Psi^{(k,l)} \right] \tag{5.2}$$

The validation loss of the accelerometer model (Figure 5-2) :

$$\mathcal{L}^{(k)} = \mathrm{AM}\left(\mathbf{X}^{(k)};\ \mathcal{A},\ \boldsymbol{\beta}^{(k)} \right) \tag{5.3}$$

was a function of the data subsample, $\mathbf{X}^{(k)}$, for *k*-th outer fold, which was conditional on the chosen algorithm, $\mathcal{A}$, and the parameters, $\boldsymbol{\beta}^{(k)}$, which for this chapter comprised the algorithm's hyperparameters (Table 5-1). The AM incorporated data preprocessing, FPCA, model training and model evaluation inside the inner CV loop. For a given algorithm, $\mathcal{A}$, and parameter set, $\boldsymbol{\beta}^{(k)}$, the validation loss, $\mathcal{L}^{(k)}$, varied depending on the CV random splits defining for $\mathbf{X}_{+}^{(k,l)}$ and $\mathbf{X}_{-}^{(k,l)}$. The aim was to determine a suitable $\hat{\boldsymbol{\beta}}^{(k)}$ for the *k*-th subsample, $\mathbf{X}^{(k)}$, that minimised the outer validation loss:

$$\hat{\boldsymbol{\beta}}^{(k)} = \arg\min \left[\mathrm{AM}\left(\mathbf{X}^{(k)};\ \mathcal{A},\ \boldsymbol{\beta}^{(k)} \right) \right] \tag{5.4}$$

5.2.7 Optimisation

The AM was computationally expensive to evaluate, primarily due to the multiple times FPCA and model training was performed in the CV loop. The AM also exhibited stochastic behaviour due to the random subsampling, which would appear as noise to an optimiser. Bayesian optimisation is well suited to finding the global minimum of such an objective function as the optimiser requires fewer function evaluations than other procedures (Brochu et al., 2010; D. Jones et al., 1998; Snoek et al., 2012). However, a preliminary investigation established that the Matlab Bayesian optimiser could not be relied upon to find the global optimum consistently (Appendix E.4). Even though Bayesian optimisation is specifically designed to handle such 'noisy' objective functions, inconsistent results is a known issue that remains the subject of research (Gramacy et al., 2016; Gramacy & Lee, 2011; Letham et al., 2019; Picheny, Ginsbourger, et al., 2013; Picheny, Wagner, et al., 2013).

Bayesian optimisation works by constructing a surrogate model from the observations made of the objective function. The more observations it gathers, the more representative the surrogate model becomes. However, the number of observations that the Bayesian optimiser can realistically process is limited by the rising exponential cost associated with its acquisition function that directs the search. Preliminary investigations using the Matlab R2020a implementation confirmed that the optimiser overhead became the dominant factor, many times the original AM cost (Appendix E.5).

The novel approach taken in this research was to train a separate surrogate model (SM) using observations gathered from a random search (Bergstra & Bengio, 2012). Particle Swarm Optimisation (PSO) was employed to find the SM global minimum (Escalante et al., 2010; J. Kennedy & Eberhart, 1995). Thus, without the overhead of an acquisition function, it was possible to gather more observations to build a comprehensive map of the search space. The random search was progressively constrained using a mechanism described below (Section 5.2.9) to concentrate observations in promising regions of the parameter space where SM loss predictions were comparatively low. PSO was configured with 100 particles to find the global optimum in the SM. The convergence criterion was

when the best objective function observed across the swarm did not improve by 0.01 $W \cdot kg^{-1}$ over 20 consecutive iterations. Typically 10-30 iterations were required, and so the number of SM function evaluations could often exceed a thousand. Given the SM's fast evaluation time (SM < 0.001 s vs. AM ~ 1.3 s), this was not an issue, and the optimisation was typically completed in a few seconds.

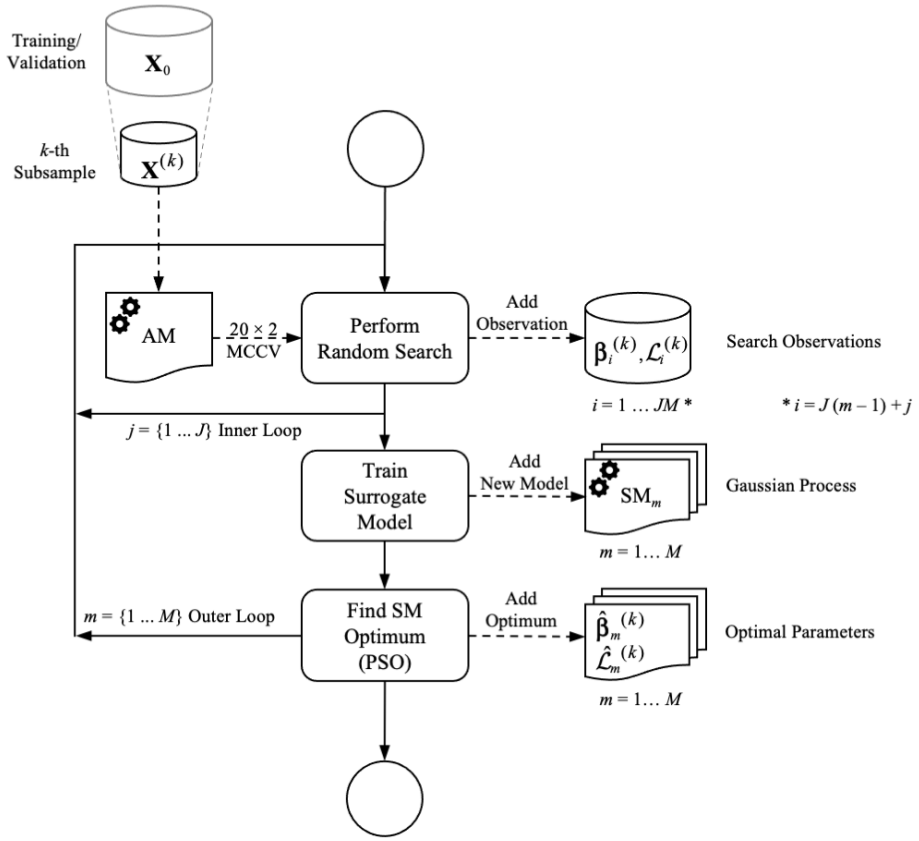


Figure 5-3. Optimisation procedure for model selection based using a random search to develop a surrogate model (SM) for predicting the validation loss, $\hat{\mathcal{L}}_m^{(k)}$ for the accelerometer model (AM) at the *m*-th step, using the optimal parameters, $\hat{\beta}_m^{(k)}$.

5.2.8 Surrogate model

The SM took the same form as the surrogate model used by the Matlab Bayesian optimiser, which was based on a Gaussian Process (GP) (**fitrgp** Matlab function). The GP used a constant basis function, no prior standardisation of the parameters and an anisotropic Automatic Relevance Determination (ARD) Matérn 5/2 kernel. The noise level parameter, σ, was fitted to the data rather than being held constant. Thus, as a GP posterior function

with adaptive noise fitting, it could model stochastic functions such as the AM validation loss.

For each *k*-fold, the SM was trained on the AM validation loss observations, $\mathbf{Z}^{(k)}$, comprising the parameters employed and the resultant loss:

$$\mathbf{Z}^{(k)} = \left[\boldsymbol{\beta}_i^{(k)} \ \mathcal{L}_i^{(k)} \right]^{\mathrm{T}}, \ \mathrm{i} \in \{1, \ldots, I\}, \ \boldsymbol{\beta}_i^{(k)} \in \mathbb{R} \tag{5.5}$$

where for this chapter, $I = 500$ observations were gathered for each fold and model. The aim of the optimisation was re-defined as finding the global minimum of the SM:

$$\hat{\boldsymbol{\beta}}^{(k)} = \arg\min \left[\mathrm{SM}\left(\boldsymbol{\beta}^{(k)} \right) \right] \tag{5.6}$$

The resulting optimal hyperparameters, $\hat{\boldsymbol{\beta}}^{(k)}$, could then be applied to the original AM, Equation (5.3).

The parameter vector, $\boldsymbol{\beta}_i^{(k)}$, was composed of purely numeric elements generated at random according to constraints defined below (Section 5.2.9), so the SM would be compatible with PSO, which operates with real variables (Table 5-3). The categorical parameters were assigned integer values (Unler & Murat, 2010) and explicitly identified as such to the SM so they would be treated as nominal rather than ordinal parameters ('CategoricalPredictors' name-value pair for the fitrgp function in Matlab). Numerical parameters spanning several orders of magnitude were log-transformed. The SM could not be made the objective function directly because PSO passed arguments as real numbers. Hence, an intermediate function served as the objective function, which rounded the values for categorical and integer parameters to the nearest whole numbers (Figure 5-4). The SM needed to appear to the optimiser as a stepped function for these variables across the PSO variables' continuous range. For the PSO search, the lower bounds of the categorical and integer parameters were set at 0.50, while the respective upper bounds were 0.49 above the highest integer value (Table 5-3, final column). These boundaries ensured each parameter value had equal 'width' in the PSO continuum.

Table 5-3. Hyperparameter ranges defining the random search space for the shortlisted accelerometer models (AM) and the corresponding initial PSO search ranges.

Model	Parameter	Type †	Hyperparameter Values (AM)	PSO Initial Bounds (SM) ‡
LR	Regularisation	C	{Ridge, Lasso}	[0.50, 2.49]
	Solver	C	{SVM, Least Squares}	[0.50, 2.49]
	Lambda	R	10^{-12} … 10^{12}	[-12, 12]
SVM	Kernel	C	{Gaussian, Linear, Polynomial} [a]	[0.50, 3.49]
	Box Constraint	R	10^{-6} … 10^{8}	[-6, 8]
	Kernel Scale	R	10^{-6} … 10^{8}	[-6, 8]
	Epsilon	R	10^{-4} … 10^{3}	[-4, 3]
	Standardisation §	C	{No, Yes}	[0.50, 2.49]
GPR	Basis	C	{None, Constant, Linear, Pure Quadratic}	[0.50, 4.49]
	Kernel	C	{Squared Exponential, Exponential, Matérn 3/2, Matérn 5/2, Rational Quadratic}	[0.50, 5.49]
	Sigma	R	10^{-4} … 10^{2}	[-4, 2]
	Standardisation §	C	{No, Yes}	[0.50, 2.49]

† Parameter type: C = Categorical (nominal variable); R = Real (continuous).
‡ Categorical parameters are indexed; real parameters are log-transformed. See text.
§ Standardise the predictors and outcome variables as Z scores during fitting.
[a] Polynomial kernel was second order

Figure 5-4. Data flow diagram illustrating how the objective function is intermediate between PSO and the SM, rounding the categorical and integer parameters

5.2.9 Random search

The SM was retrained every 20 iterations of the random search, producing a series of models, $\mathrm{SM}_m^{(k)}$, where $m \in \{1 \ldots M\}$ (Figure 5-5). In this chapter, $M = 25$. The majority of observations (95%) came from the random search, but every 20th observation was positioned at the SM_{m-1} global optimum, located by PSO on the preceding iteration. In this

way, the SM learned the accuracy of its global minimum estimate, which improved the optimisation efficiency.

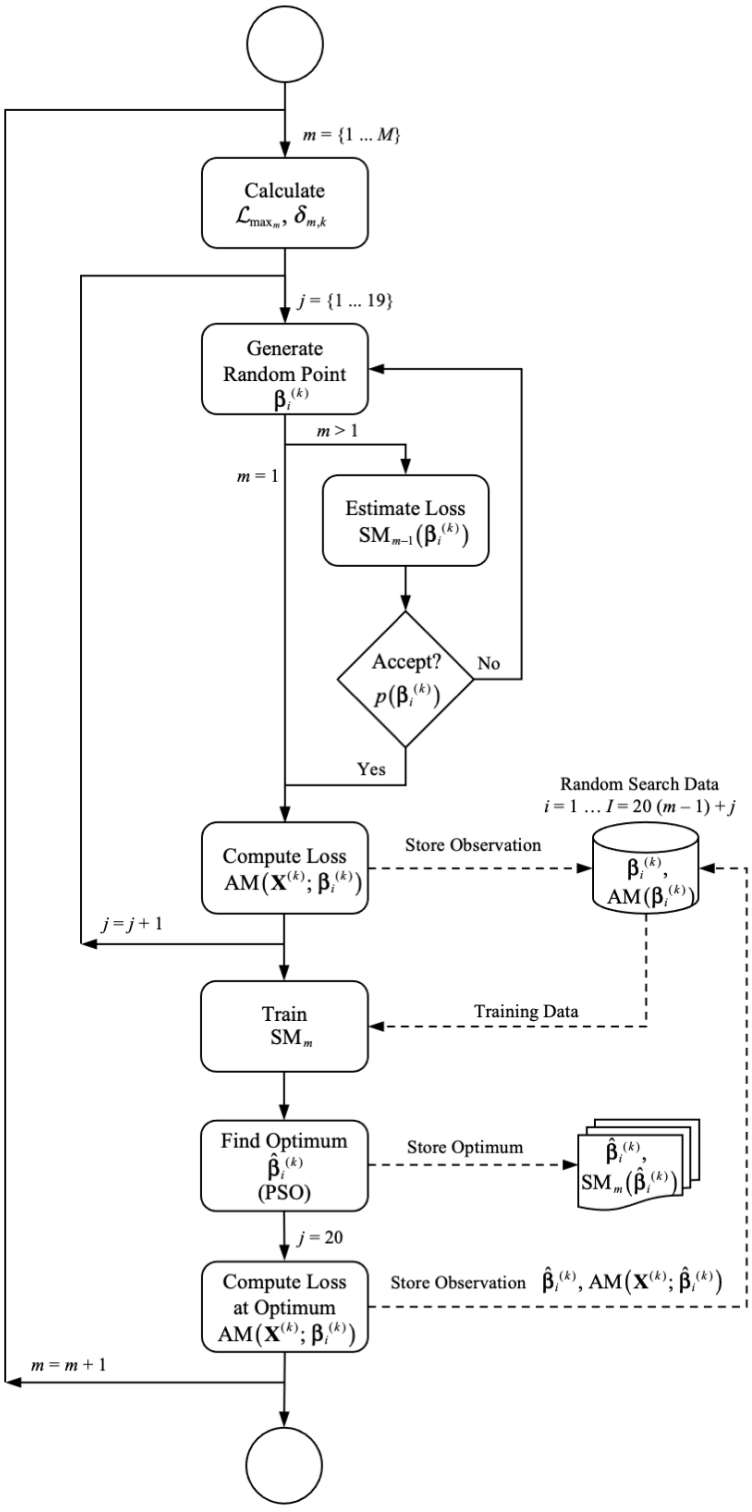


Figure 5-5. Random search process flow. See text for definitions.

The random search was constrained to progressively concentrate observations in promising regions as follows. Each generated random point, $\boldsymbol{\beta}_i^{(k)}$, was within the range specified in Table 5-3, following the conventional random search algorithm. However, this point was only accepted according to the probability function:

$$p\left(\boldsymbol{\beta}_i^{(k)}\right) = \begin{cases} 1, & \mathcal{L}_i \le \mathcal{L}_{\max_m} \\ \exp\left[\dfrac{-\left(\mathcal{L}_i - \mathcal{L}_{\max_m}\right)^2}{2\delta_{m,k}^{\,2}}\right], & \mathcal{L}_i > \mathcal{L}_{\max_m} \end{cases} \tag{5.7}$$

which depended on the SM loss prediction at that point, $\mathcal{L}_i = \mathrm{SM}_{m-1}\left(\boldsymbol{\beta}_i^{(k)}\right)$. $\mathcal{L}_{\max_m}$ is the maximum loss, defined below (Equation (5.10)), and $\delta_{m,k}$ is the standard deviation in the re-scaled normal distribution, represented by the exponent function above (Figure 5-6). Hence, if the prediction, $\mathcal{L}_i$, was less than the maximum loss, $\mathcal{L}_{\max_m}$, $\boldsymbol{\beta}_i^{(k)}$ was always accepted. If not, the chances of $\boldsymbol{\beta}_i^{(k)}$ being accepted diminished the higher $\hat{\mathcal{L}}_i$ was above $\mathcal{L}_{\max_m}$. The probability function, $p\left(\boldsymbol{\beta}_i^{(k)}\right)$, could be evaluated quickly because it depended on $\mathrm{SM}_{m\text{-}1}$ rather than AM, allowing potentially large numbers of points to rejected before one was accepted. The accepted point was then used to generate a new AM loss observation.

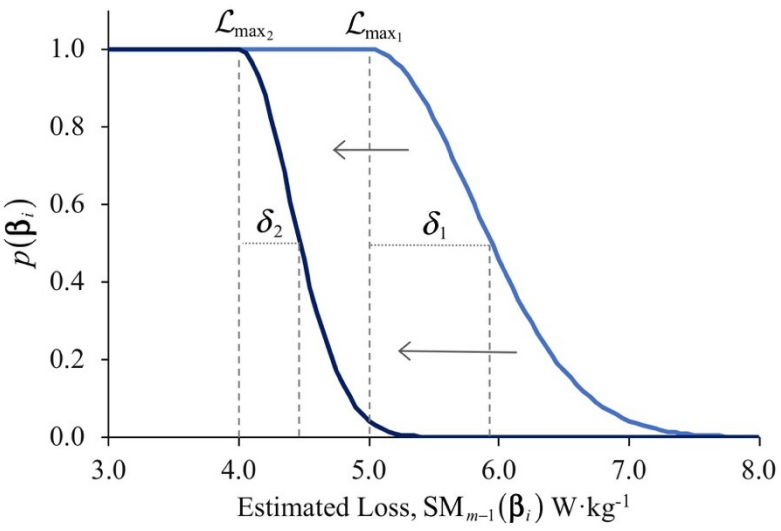


Figure 5-6. The function defining the probability of a randomly generated point being accepted based on the estimated loss at that position. The progressive reduction in the maximum loss is illustrated by the transition from the light blue line to the dark blue line ($\mathcal{L}_{\max_1} \rightarrow \mathcal{L}_{\max_2}$), with a narrowing in the standard deviation ($\delta_1 \rightarrow \delta_2$).

The advantage of a stochastic upper limit was that it allowed observations outside of the concentrated regions to be included occasionally in case the true SM minimum happened to lie elsewhere. This stochastic upper limit had the same purpose as the exploration function in Bayesian optimisation that ensures the optimiser does not overexploit a region. The drop off in probability was governed by, $\delta_{m,k}$ which was defined as half the standard deviation in AM observations up to that point:

$$\delta_{m,k} = \frac{1}{2}\sqrt{\frac{\sum_{i=1}^{20(m-1)}\left(\mathcal{L}_i^{(k)} - \overline{\mathcal{L}}_{m-1}^{(k)}\right)^2}{20(m-1)}} \tag{5.8}$$

where the mean AM loss was given by

$$\overline{\mathcal{L}}_m^{(k)} = \frac{1}{20m}\sum_{i=1}^{20m} \mathrm{AM}\left(\mathbf{X}^{(k)};\, \mathcal{A},\, \boldsymbol{\beta}_i^{(k)}\right) \tag{5.9}$$

The factor of ½ was introduced for convenience to more tightly constrain the stochastic region. The number of observations was $20(m-1)$ rather than $20m$ to reflect the fact that $\delta_{m,k}$ must be determined from the previous $m-1$ iterations. This definition yielded a standard deviation that was well-suited to its purpose as it gradually decreased as the search progressed, sharpening the observation cut-off as $\mathcal{L}_{\max_m}$ approached the SM minimum (Figure 5-6). The SD had a typical range of 0.3 to 0.8 W·kg^{-1}.

The maximum loss was defined as a function of the *m*-th iteration:

$$\mathcal{L}_{\max_m} = \begin{cases} \mathcal{L}_{\max_0}, & m = 1 \\ \left(2 - \dfrac{m}{M}\right)\arg\min_{\mathrm{PSO}}\left[\mathrm{SM}_{m-1}(\boldsymbol{\beta}_i)\right], & m > 1 \end{cases} \tag{5.10}$$

where m/M is the proportion of the search completed so far. Except for the first iteration, $\mathcal{L}_{\max_m}$ started at nearly twice the first SM global minimum, determined by PSO, and dropped in steps until it reached one on the final iteration. In the first iteration, when there is no trained SM, the maximum loss was set to a user-defined large number, defined in this study as 10 W·kg^{-1}, which was intended to encompass almost all observations. A limit was imposed to avoid instability should the AM produce a large loss (e.g. very occasionally for

SVM $\mathrm{AM}(\boldsymbol{\beta}_i) > 10^3$ W·kg^{-1}), which distorted the SM landscape and led to erratic predictions.

5.2.10 Bagging

An optimisation yielded a series of optimal parameter values, $\hat{\boldsymbol{\beta}}_m^{(k)}$, accompanied by estimates of the inner and outer AM loss, the latter defined as $\mathcal{L}_m^{(k)}$ (Figure 5-7A). There were *k* separate model series from which a single aggregate model was desired, which by definition would be representative of a broader set of data. However, a given model series, $\hat{\boldsymbol{\beta}}_m^{(k)}$, did not converge to a stable vector, as might be expected. Hence, there was no single final optimal model. Therefore, it was appropriate for the model aggregation to encompass a range of models $\mathrm{SM}_m^{(k)}$ from a given series (varying *m*), as well as models across series (varying *k*), that is, within and between outer folds (Figure 5-7A).

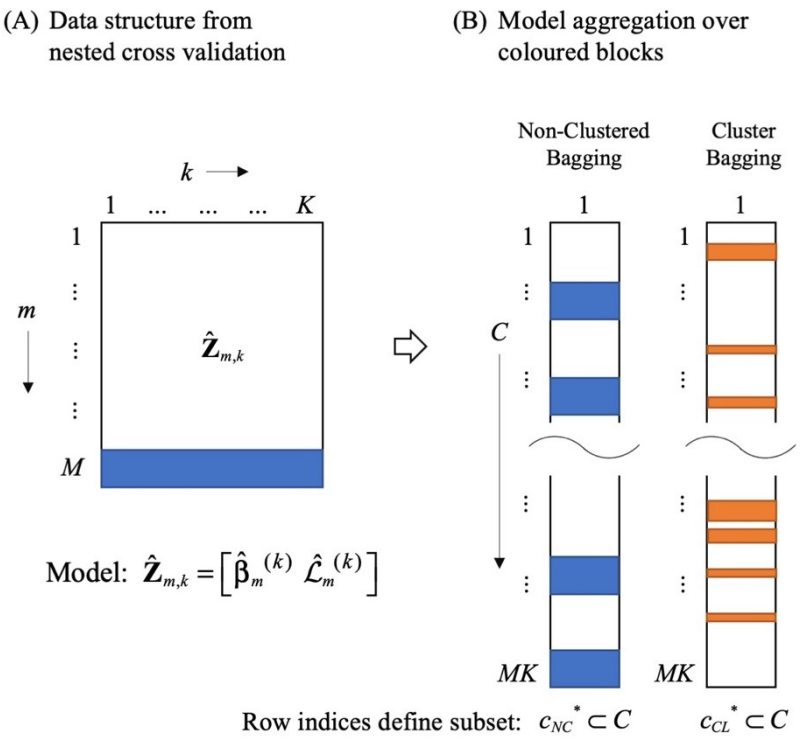


Figure 5-7. Data structure for bagging. (A) Nested cross validation yields a set of optimal models represented by a *k* × *m* matrix $\hat{\mathbf{Z}}$, where each element includes the parameter vector $\hat{\boldsymbol{\beta}}$ and the outer AM loss $\hat{\mathcal{L}}$. Each matrix column represents a series of models of length *m*, and each row is a cross-section of the models across the *k* outer folds. (B) $\hat{\mathbf{Z}}$ is reshaped to a single column of *MK* rows, which is indexed by *C*. Non-clustered bagging aggregates models from repeating fixed blocks defined by $m \bmod m_0$ for all *k*. Clustered bagging aggregates models with similar $\hat{\boldsymbol{\beta}}$, which results in multiple index sets, $\mathbf{c}_{\mathbf{CL}}^*$

The aggregation method adopted was bagging (Breiman, 1996), where selected models were given equal weighting. For the bagged model, the aggregate values, $\hat{\boldsymbol{\beta}}_{\mathrm{BAG}}$, for the numeric parameters and losses were defined as the means of those quantities over the chosen model subset. The bagged categorical parameters were those with the highest frequency. The SM's predictions were initially poor and needed to improve before those models could be included. The criterion was for the mean absolute error in the SM's prediction of the AM inner loss

$$\mathrm{MAE}_m^{(k)} = \left| \mathrm{SM}_m^{(k)} - \mathrm{AM}_m^{(k)} \right| \tag{5.11}$$

to have stabilised to a comparatively low level, which was determined subjectively by inspection. The bagged model was not particularly sensitive to this choice, as results from other reasonably chosen start points made clear. The start point, m_0, defined the lower bound of the model series, such that $m \geq m_0$ for all *k*, a model subset referred to as NCA (non-clustered, all models). On the assumption that the parameters later in the series would be closer to the true optimum values, by virtue of the $\mathrm{SM}_m^{(k)}$ being trained on a larger data set, a second non-clustered model subset was defined as $m \geq M - 5$, called NC5, so that only the final five models in each series were included.

It was apparent that certain distinct values were favoured for each component of $\hat{\boldsymbol{\beta}}_m^{(k)}$ rather than those elements having a continuous distribution. This outcome pointed towards an alternative strategy of defining the model subset based on a cluster analysis undertaken in SAS 3.8, where $m \geq m_0$. The clusters were obtained using SAS PROC CLUSTER, a hierarchical clustering procedure where the centroid method was chosen. The proximity between the models was calculated using PROC DISTANCE, based on the Gower distance that is suitable for positions defined by both nominal (categorical) and ratio (numerical) variables (Gower & Legendre, 1986). Before calculating the Gower distance, the categorical variables were transformed into their mid-rank score, which took account of their frequency, before all variables were standardised to [0, 1] based on their range. Peak values in the pseudo t^2-statistic determined the number of clusters. When there was more than one prominent peak, the preference was for fewer, larger clusters across the outer folds. The five largest clusters were retained for analysis, designated by CL*n*.

5.2.11 Model Evaluation

The final models from nested cross validation were retrained on $\mathcal{D}_0$ using their optimal parameters, $\hat{\boldsymbol{\beta}}_{\mathrm{BAG}}$ (Varma & Simon, 2006), to determine their predictive accuracy with larger data sets than were used in the optimisation (Figure 5-8). In the first instance, the models were evaluated using a ten-fold design for a low bias predictive error (Y. Zhang & Yang, 2015) in which the training sets were 90% of $\mathcal{D}_0$. The evaluation was based on 1000 repetitions of MCCV to minimise data subsampling variation (Xu et al., 2004; Xu & Liang, 2001). The estimate was equivalent to the average reported error from 1000 researchers independently gathering and processing the data with identical procedures. Secondly, the models were trained on the entire $\mathcal{D}_0$ and then tested on the holdout test data set, $\mathcal{D}_{\mathrm{H}}$. Since the holdout data set was kept entirely separate from the model selection procedures, it was a truly independent test (Kohavi, 1995; Vabalas et al., 2019, 2020).

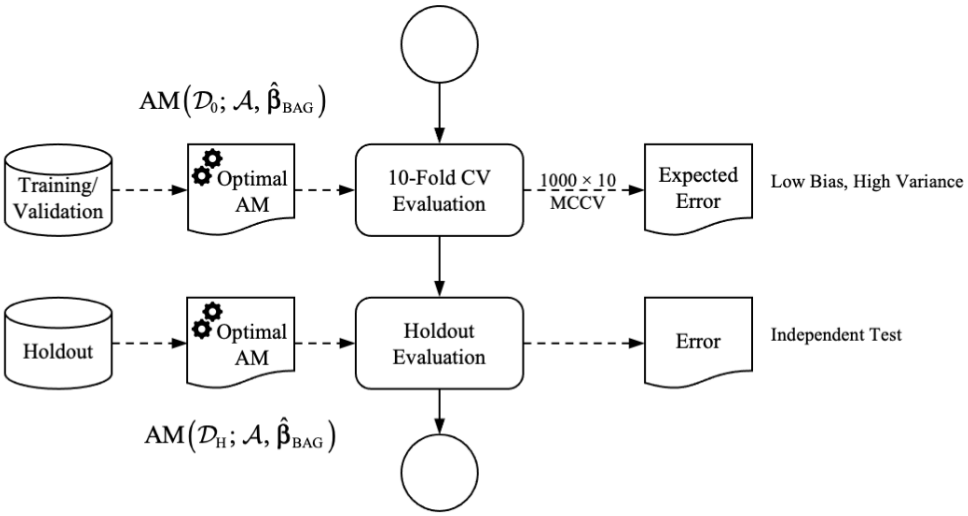


Figure 5-8. Process flow for model evaluation showing three different ways of doing so once the optimal accelerometer models (AMs) have been identified. MCCV = Monte Carlo Cross Validation.

5.3 Results

5.3.1 Initial Model Selection

The GLM, which compared 4,000 accelerometer model evaluations across all grid searches (2 jump types × 4 sensors × 10 algorithms × 50 iterations), had a total explained variance, $\omega^2 = 0.98$ and RMSE = 0.21 W·kg^{-1}. More than half the variance in AM predictive error was explained by the sensor (55.6%), followed by the jump type (36.7%) and the algorithm (6.9%) with the remaining explained proportion (7.7%) explained by interaction effects (Table 5-4).

Table 5-4. Overall explained variance of jump type, sensor attachment site and algorithm on the accelerometer models' accuracy.

Effect	DF †	*F*-Statistic ‡	Explained variance §
Sensor	3	38,074	0.533
Jump Type	1	72,545	0.338
Algorithm	9	1,237	0.052
Sensor × Jump Type	3	2,923	0.041
Sensor × Algorithm	27	86	0.011
Jump Type × Algorithm	9	37	0.002
Sensor × Jump Type × Algorithm	27	40	0.005

† Degrees of Freedom: Model Total = 79, Error = 3920, Corrected Total = 3999
‡ $p < 0.0001$ in all cases. Overall model *F*-statistic = 2,663, $p < 0.0001$.
§ Semi-partial ω^2

The LB-sensor models had a smaller RMSE than the UB-sensor models, on average across jump types and all algorithms, 5.20 W·kg^{-1} versus 6.12 W·kg^{-1} (Table 5-5). The difference between LS and RS models, where the sensors were placed at the same position on each leg, was 0.33 W·kg^{-1} on average. Errors for the CMJ_{NA} were lower than for the CMJ_A, 5.83 W·kg^{-1} versus 7.60 W·kg^{-1}, when averaged across all sensors and algorithms. The combination of LB sensor and CMJ_{NA} achieved the lowest predictive error of all, where RMSE = 3.97 W·kg^{-1}, averaged across all algorithms.

Table 5-5. Accelerometer model predictive error for different jump types and sensors, based on a grid search using the hyperparameter values set by the Matlab HPO. Least mean squares estimates shown, averaged across all algorithms. All units W·kg^{-1}.

Sensor	CMJ_{NA}	CMJ_{A}	Overall †
LB	3.97	6.43	5.20
UB	4.96	7.27	6.12
LS	7.36	8.50	7.93
RS	7.01	8.19	7.60
Overall ‡	**5.83 (13.0%)**	**7.60 (14.8%)**	**6.71 (13.9%)**

† Overall effect, averaged across jump type, according to the GLM.

‡ Overall effect, averaged across all sensors, according to the GLM. Percentage error calculated by dividing the error by the mean peak power across the training/validation data set for the jump type. This calculation was post-hoc as the alternative of dividing individual peak power predictions by the criterion value would have altered the loss function, requiring a separate cross-validation.

Grouping the models by algorithm type for CMJ_{NA}, the SVM models were the best for the LB, LS and RS sensors, followed by LR, GPR, NN and TR, in descending order (Table 5-6). For the UB sensor, LR achieved the lowest error, followed by SVM, but otherwise, it followed the same ranking order. At the level of the individual algorithm, the SVM-L model reported the lowest error for each sensor, and for the LB sensor, it achieved the lowest error of all with an RMSE = 3.40 W·kg^{-1}. Generally speaking, the SVM and LR models were consistently ranked among the best with little difference between them, whilst in a clear third place were the GPR models. The pattern of results for individual models was similar across the sensors, as illustrated in Figure 5-9 for LB and UB sensors for CMJ_{NA}. This consistency reflected the top-level analysis from Table 5-4, which showed that algorithm was a relatively small factor compared to sensor and jump types. It follows then that the results from the LB+CMJ_{NA} data set will generally be representative of those from other combinations of sensor and jump type. Therefore, the LB sensor data set for the CMJ_{NA} was selected for further analysis as it had the lowest error. Similarly, the SVM, LR and GPR models were shortlisted as their errors were the lowest.

Table 5-6. Accelerometer model predictive error for different algorithms for CMJ_{NA}, grouped by algorithm type, for different sensors.

Type	Algorithm	LB	UB	LS	RS	All Sensors
LR	LR	3.69	4.60	7.45	6.97	5.68
	LR-RDG	3.65	4.60	7.36	6.93	5.64
	LR-LSS	3.49	4.50	7.29	6.96	5.56
	Mean	**3.61**	**4.57**	**7.37**	**6.96**	**5.63**
SVM	SVM-L	3.40	4.45	6.84	6.62	5.33
	SVM-G	3.51	4.96	7.24	6.91	5.65
	Mean	**3.46**	**4.70**	**7.04**	**6.76**	**5.49**
GPR	GPR-SE	4.01	5.26	7.18	6.84	5.82
	GPR-M52	3.97	5.04	7.08	6.78	5.72
	Mean	**3.99**	**5.15**	**7.13**	**6.81**	**5.77**
NN	NN-5	4.35	5.16	7.45	7.23	6.05
	NN-10	4.94	5.73	8.17	7.94	6.70
	Mean	**4.64**	**5.45**	**7.81**	**7.59**	**6.37**
TREE	TR-ENS	**4.71**	**5.33**	**7.57**	**6.95**	**6.14**
All Algorithms		**3.97**	**4.96**	**7.36**	**7.01**	**5.83**

Least mean squares estimates shown ($W \cdot kg^{-1}$).
Emboldened figures are mean values for the algorithm type or for all algorithms at the bottom of the table.

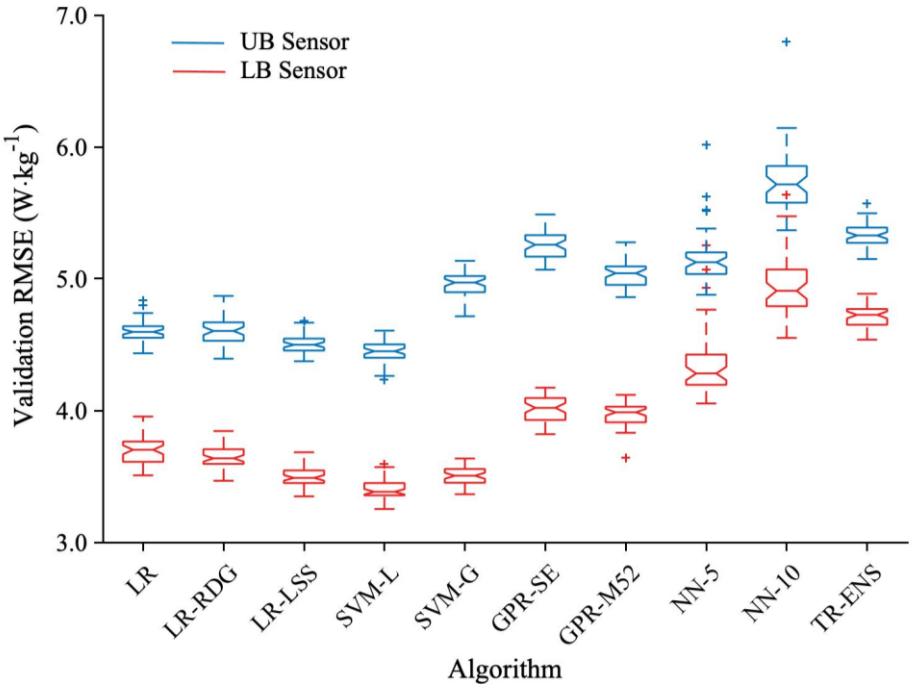


Figure 5-9. Predictive error for the top two accelerometer model algorithms for CMJ_{NA}: Lower Back, LB (red) and Upper Back, UB (blue). Note the similar pattern of UB and LB results, rising and falling together.

5.3.2 Optimisation of shortlisted models

The progression of the optimisation procedures, $m = 1 \ldots M$, for the LR, SVM and GPR models are presented in Figure 5-10, which aggregates the quantities of interest over the ten outer folds. The mean absolute error (MAE) in the SM predictions of the inner AM validation loss stabilised at a lower value more quickly for the LR models than for the SVM and GPR models, as indicated by the lightly shaded regions (Figure 5-10A–C). Toward the end of the optimisation, MAE was 0.07 $W \cdot kg^{-1}$, 0.37 $W \cdot kg^{-1}$ and 0.09 $W \cdot kg^{-1}$, respectively, for the three model types, when averaged over the final five models ($m \geq 20$), equivalent to the final 100 iterations. The five-model moving correlation (r^2) between the inner and outer AM validation losses followed a downward trend for all models to a minimum of 0.07–0.16 $W \cdot kg^{-1}$, indicating the inner and outer losses effectively ceased to have any relation with one another. However, the trend appeared to reverse in the final 100 iterations or so, notably for the GPR model.

Reflecting the accuracy in the SM's MAE, the fitted noise level and the confidence interval was lowest for the LR model, followed by the GPR and then the SVM models (Figure 5-10D–F). The SM confidence level at the global optimum converged towards the noise level, reflecting the concentration of observations in that region. As the optimisation progressed, noise became the dominant factor over model uncertainty. MAE was higher for the SVM because of the higher level of uncertainty that arose from the greater noise level. In contrast, the SM could predict the LR AM losses with a higher level of accuracy because the SM's confidence level almost merged with the low noise.

The SM prediction errors were averaged out in the aggregated losses, which stabilised to broadly constant values for SM and the inner and outer AMs (Figure 5-10G–I). The SM loss prediction had a small variance (0.10–0.13 $W \cdot kg^{-1}$ for the three model types averaged over the final five models), and the inner AM loss had more variance (0.13–0.14 $W \cdot kg^{-1}$). The largest variance of all was seen in the outer AM loss (0.66–0.79 $W \cdot kg^{-1}$), which represented the independent test of the model. Notably, in all three models, the aggregate outer AM loss was lower than the aggregate inner AM loss, reflecting the bigger data set used to train the outer model compared to the inner set.

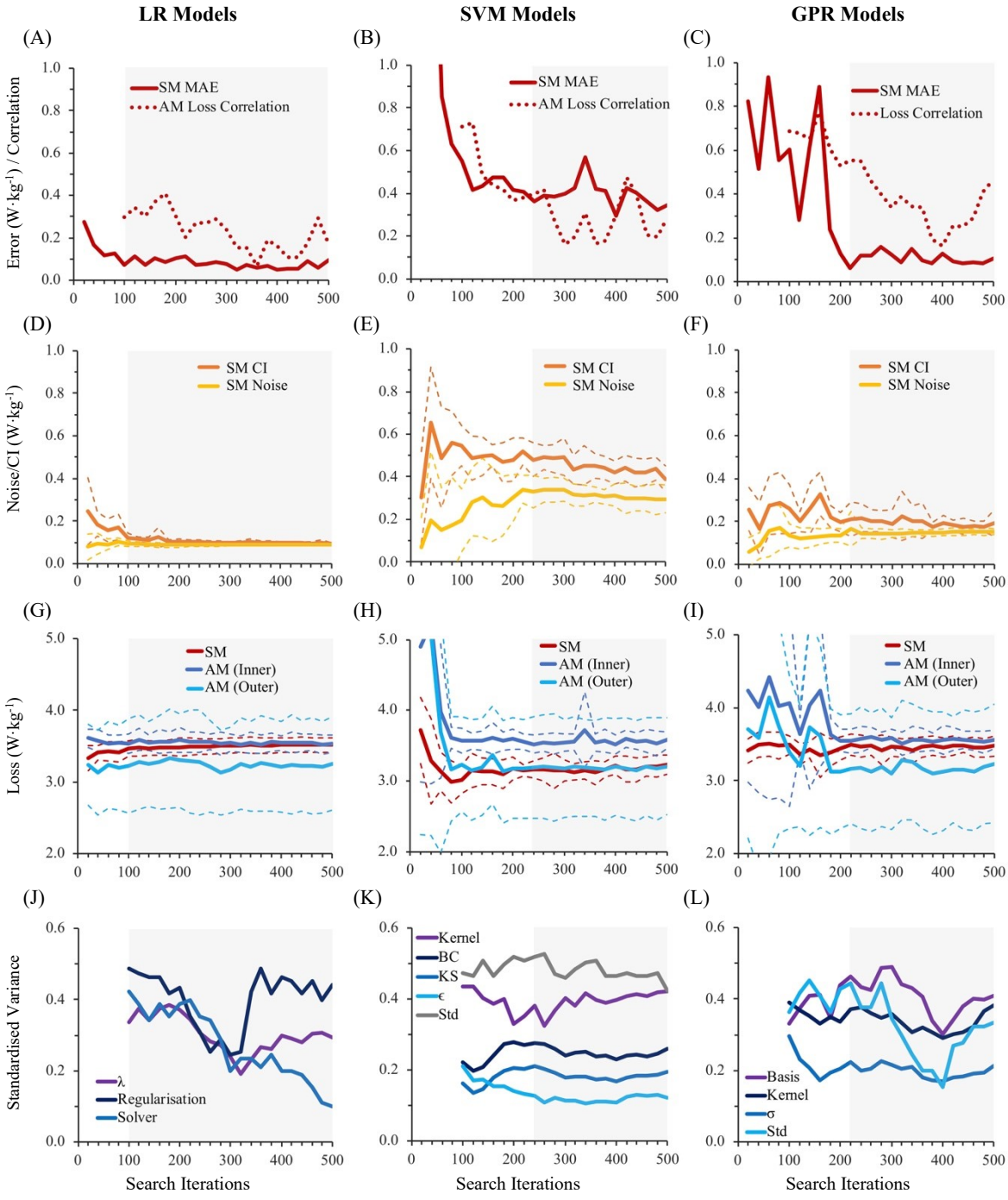


Figure 5-10. Optimisations for LR, SVM and GPR models, averaged over ten outer folds: means (solid lines), SDs (dashed lines). (A, B, C) Absolute error in SM prediction of AM inner loss (red line); 5-point moving correlation (r^2) of AM inner and outer losses (red dots). (D, E, F) SM confidence interval (orange), expressed as SD, converges to noise level (yellow). (G, H, I) SM loss prediction (red), actual AM inner loss (blue) and independent AM outer loss (light blue). (J, K, L) SM parameter variances (numeric form) based on 5-point moving SD, standardised to parameter range. BC = Box Constraint, KS = Kernel Scale, Std = Standardisation. Lightly shaded area indicates the range used for cluster analysis.

The five-point moving average of parameter variance in aggregate was relatively low for some parameters (e.g. ε for SVM and σ GPR), indicating that those parameters had stabilised to more or less constant values (Figure 5-10J–L). In the case of the LR solver, the parameter values continued to drop throughout. However, the variance for other parameters remained moderate or high (e.g. standardisation and kernel for SVM, basis and kernel for GPR). There were others for which variance dropped but then rose (regularisation for LR and standardisation for GPR). The reversal of the trend towards rising parameter variance indicated a growing sensitivity of the SM to more AM observations, which meant the global optimum was jumping between different but similarly valued minima in the parameter space.

The clustering procedure was run on the parameters from the subset of SM_m series starting from 100 iterations for LR ($m \geq 5$), 240 iterations for SVM ($m \geq 12$) and 220 iterations for GPR ($m \geq 11$), when MAE stabilised, as indicated by the lightly shaded regions in Figure 5-10. Based on the cluster selection procedure (Section 5.2.10), 28, 17 and 30 clusters were chosen respectively for the LR, SVM and GPR models. Details on the five largest clusters for each model are shown in Figure 5-11. The largest cluster for LR (CL3) with 72 models stood out from others, whilst the SVM and GPR clusters were of a similar size, the largest being CL1 for SVM and CL2 for GPR, both comprising 26 models each (Figure 5-11A–C). The spread of the top-five clusters across the outer folds was irregular such that some clusters could dominate some folds while being almost absent in others. Such an uneven spread suggested subsampling variation could substantially alter the optimal model (Figure 5-11D–F). The biggest clusters were therefore the ones that were the least sensitive to variations in the data distribution. The clusters' appearance in the SM_m series generally reflected the prescribed m ranges above, indicating that similar models could recur many times as more observations were added (Figure 5-11G–I). Hence, there was no obvious point at which the optimisation should stop based on the SM_m series.

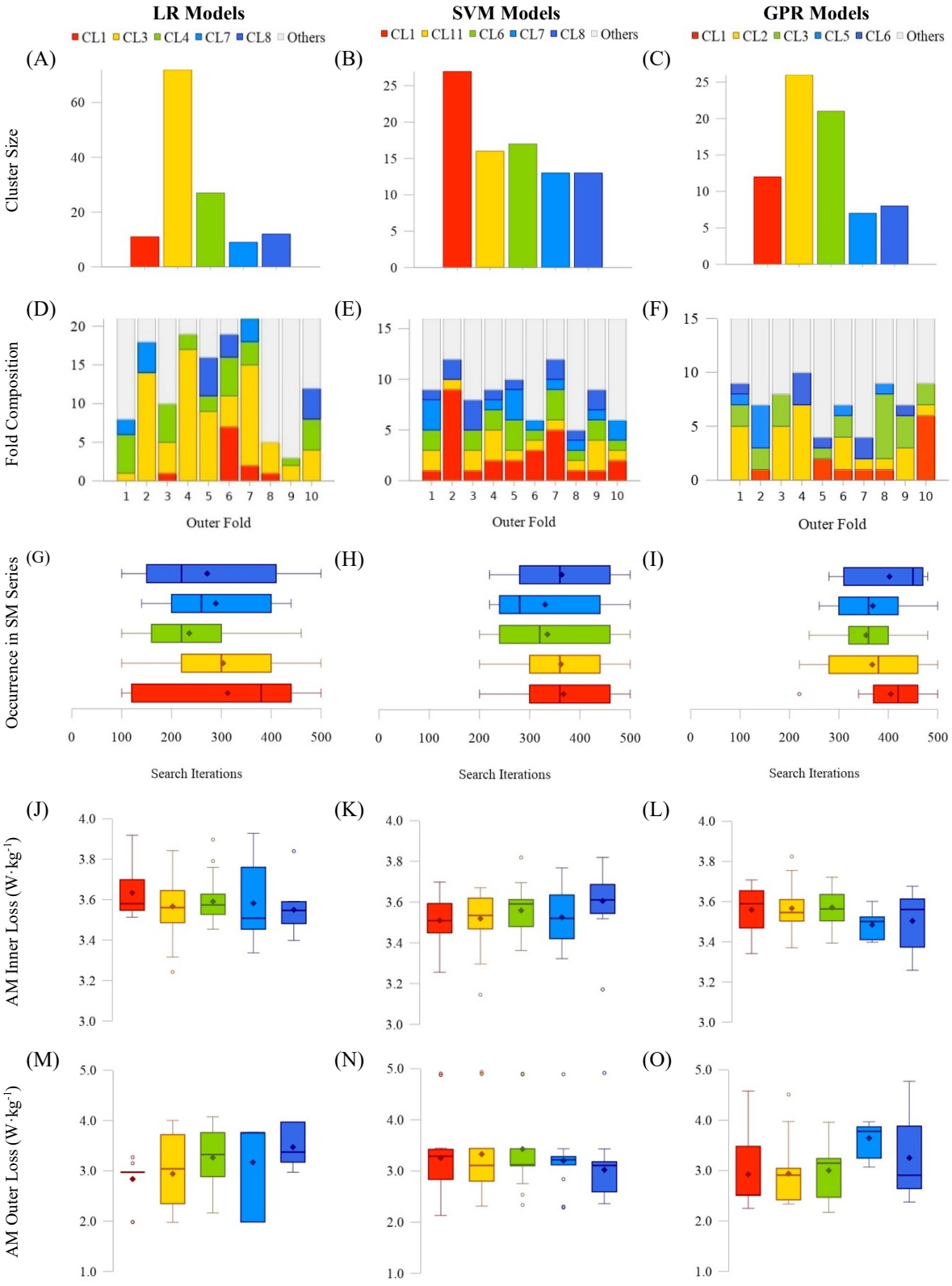


Figure 5-11. Clusters (five largest) obtained from SM optimisation series, identified and colour-coded at the top per model type. (A, B, C) Prevalence with which the clusters appeared in the outer folds; other smaller clusters in grey. (D, E, F) Cluster sizes – number of models summed across outer folds. (G, H, I) Preponderance of times when the clusters were observed in the SM optimisation series. (J, K, L) Spread of AM loss per cluster for the inner validation set. (M, N, O) AM loss spread per cluster for the outer validation set. Box plots: diamond = mean; line = median; box = IQR; whiskers = 95% CI; circles = outliers.

The most important statistic from the optimisation was the AM outer loss as it was the independent test of the optimised model (Figure 5-11M–O). As previously noted in the aggregate plots above (Figure 5-10G–I), the outer loss was lower and with a much larger variance than for the AM inner loss (Figure 5-10J–L). However, the outer loss interquartile range differed considerably between clusters, indicating that their performance on independent test data (outer validation set) could be quite inconsistent in some cases. The LR clustered models' performance was the least reliable, exhibiting some of the largest and smallest variations (Figure 5-10G). The SVM clustered models had the smallest interquartile range, but they also registered the most outliers with several at very high losses, $\hat{\mathcal{L}}_m > 4.8\,\mathrm{W\cdot kg^{-1}}$ (Figure 5-10M). The GPR clustered models had moderate interquartile ranges with only one outlier (Figure 5-10O).

The bagged parameters are shown in Table 5-7, Table 5-8 and Table 5-9 for the clustered and non-clustered model subsets. For the LR model, ridge regularisation and the SVM solver were most frequently chosen (NCA), which also defined the largest cluster (CL3) (Table 5-7). The regularisation parameter, λ, could vary over many orders of magnitude, but it remained less than unity. Whilst the non-clustered subsets had similar λ in the mid-range, by definition, clustering produced values at specific points in the range. For SVM models, the linear kernel was by far the most common, but the polynomial kernel defined the largest cluster (Table 5-8). The polynomial kernel required a box constraint and kernel scale of a different order of magnitude compared to the linear kernels. The value of ϵ was consistent across the clusters, while the choice of whether or not to standardise the predictors varied. The non-clustered approach for the SVM model yielded the lower outer loss estimate (3.18 $\mathrm{W\cdot kg^{-1}}$). GPR models having no basis function were the most prevalent, and those with a rational quadratic kernel belonged to the largest cluster (Table 5-9). Standardisation of the predictors was by far the most frequent choice, while the value for σ was similar between clusters. The mean outer loss was lower for the clustered models suggesting that taking a subset of the models for aggregation was beneficial. For LR and GPR, three clusters achieved lower predictive errors (CL1, CL3, and CL7 for LR; CL1–3 for GPR), while there was one cluster for SVM (CL11).

Table 5-7. Clustered and non-clustered LR models.

Cluster	Size	Regularisation †	Solver †	λ_{LR} ‡	AM Outer Loss ($W \cdot kg^{-1}$) *
CL1	11	Lasso	SVM	-11.92	2.84
CL3	72	Ridge	SVM	-1.32	2.94
CL4	27	Ridge	Least Squares	-0.60	3.26
CL7	9	Lasso	SVM	-8.31	3.17
CL8	12	Lasso	SVM	-10.00	3.47
NC5	50	Ridge	SVM	-5.49	3.23
NCA	210	Ridge	SVM	-4.63	3.24

† Each cluster had a single categorical value. ‡ Mean $\log_{10}$ values shown. * Mean validation RMSE shown.
NC5 = Comprising the last 5 models from each outer fold. NCA = Comprising all models from starting point.

Table 5-8. Clustered and non-clustered SVM models.

Cluster	Size	Kernel †	Box Constraint ‡	Kernel Scale ‡	ϵ ‡	Standard-isation †	AM Outer Loss ($W \cdot kg^{-1}$) *
CL1	27	Polynomial	6.59	3.52	-1.63	N	3.25
CL6	17	Linear	-1.72	-1.13	-1.60	N	3.33
CL11	13	Linear	0.85	-0.49	-1.30	Y	3.02
CL7	13	Linear	2.61	1.19	-1.90	N	3.43
CL8	16	Linear	3.79	1.32	-2.67	Y	3.20
NC5	50	Linear	3.47	1.59	-1.96	N	3.18
NCA	150	Linear	3.86	1.72	-1.96	N	3.18

† Each cluster had a single categorical value. ‡ Mean $\log_{10}$ values shown. * Mean validation RMSE shown.

Table 5-9. Clustered and non-clustered GPR models.

Cluster	Size	Basis	Kernel †	σ ‡	Standard-isation †	AM Outer Loss ($W \cdot kg^{-1}$) *
CL1	12	None	SqExp	0.64	Y	2.92
CL2	26	None	RQ	0.56	Y	2.94
CL3	21	None	M32	0.57	Y	3.00
CL5	7	Linear	M32/M52	0.57	Y	3.64
CL6	8	Linear	M52/RQ	0.63	N	3.25
NC5	50	None	RQ	-0.14	Y	3.17
NCA	140	None	RQ	-0.01	Y	3.17

† Some clusters were split between two kernels: CL5 (3 × M32, 4 × M52); CL6 (5 × M52, 3 × RQ)
‡ Mean log10 values shown. * Mean validation RMSE shown.
SqExp = Squared Exponential; RQ = Rational Quadratic; M32 = Matérn 3/2; M52 = Matérn 5/2

Partial plots for each model revealed the performance characteristics of the models, according to the SM (Figure 5-12), where one parameter was allowed to vary while the others were held constant at their optimal values. The largest cluster for each model type was chosen to illustrate the model behaviour. From the regulation parameter plot (Figure 5-12C), λ had to have a small value, otherwise the error was large. This high level matched the peak power SD of 7.3 $W \cdot kg^{-1}$ from Chapter 3. With heavy regularisation, the model had become a naïve 'constant' model that always predicts the mean peak power whatever the input. The plot also reveals a small dip around $\log_{10} \lambda = -1.3$ where the optimum is located which explains why the NC models' underperformed because they overlooked this modest minimum by simply averaging λ across all models. The same observation can be made for the GPR where the σ plot has a similar shape but with a more pronounced dip before the sharp rise (Figure 5-12K). As with LR, by naively averaging, the NC model missed the minimum.

The SVM parameter landscape, on the other hand, does not have the nuances of the other two models as the box constraint and kernel scale loss functions are convex, while ϵ increases monotonically (Figure 5-12E–G). This behaviour would explain why the NC models perform equally well, if not better than the clustered models, because the parameters were more precisely defined by virtue of the average being taken over a larger number of models. The noise level was a major factor making it difficult for a human observer or an optimiser to discriminate between different parameters' values. The predictions for different categories were quite similar, especially for those of the SVM model (Figure 5-12B, D, H, I, J and L). It also implied that certain parameters had a weak effect on model behaviour (e.g. LR solver, GPR standardisation and the GPR kernels except for the exponential kernel). The similar loss predictions confirmed that the models were nearly equivalent for certain parameters, irrespective of the values chosen. Finally, it should also be borne in mind that the plots may look different to some extent when a different optimum is chosen. The patterns shown here reflect a localised region of the parameter space.

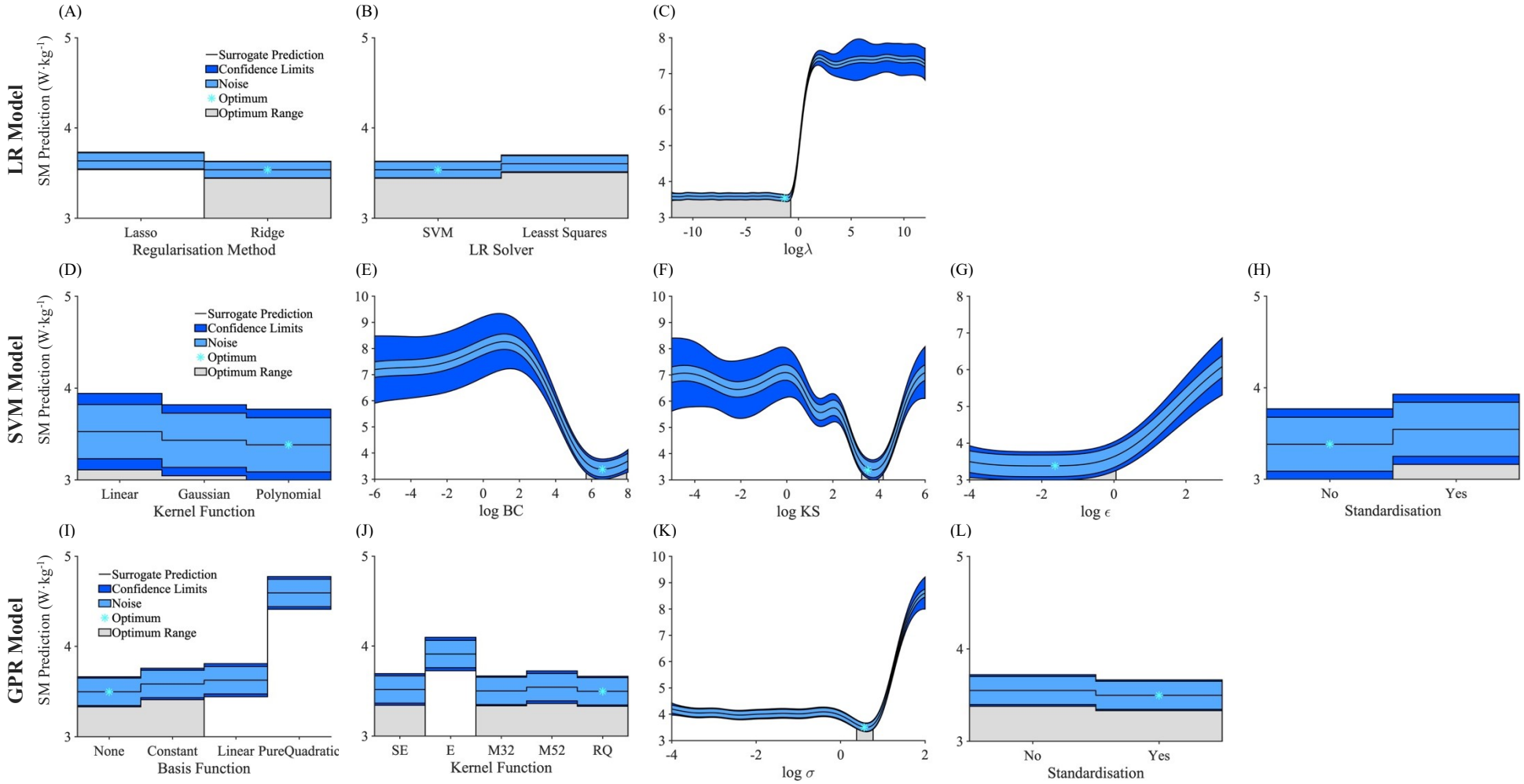


Figure 5-12. Bagged optimal SM loss prediction for the LR (CL3), SVM (CL1) and GPR (CL2) models. For each parameter, the SM prediction assumes the other parameters are held constant at their optimal values. The SM-fitted noise level is constant for each model. The SM confidence interval is the SD in the GP distribution, which tends to be large for high loss and small closer to the minima where observations were concentrated. The located optimum is based on the bagged parameter estimate, which may not necessarily be at the bagged SM prediction minimum.

The bagged parameters were then applied to models evaluated on the whole data set using 1000 × 10 MCCV for comparison with the nested figures above. The models were also trained on the full data set and then tested on the holdout data set. These three ways of estimating the generalised predictive error are presented in Table 5-10, Table 5-11 and Table 5-12. The results show that for LR, SVM and GPR models, except for a few exceptions, the NCV predictive error was lower than those of MCCV, which in turn was lower than the holdout error. The MCCV estimates, however, in most cases were within the NCV standard errors range, which was narrower for the larger clusters and the non-clustered models, as expected for a bigger sample. It was also notable that the NCV estimates varied much more between clustered models than the MCCV or holdout estimates which were relatively consistent. Finally, the residuals plots confirmed no systematic bias or heteroscedasticity was present, based on the MCCV predictions (Figure 5-13).

Table 5-10. Generalised predictive error for LR models.

Cluster	NCV Error ($W \cdot kg^{-1}$) †	MCCV Error ($W \cdot kg^{-1}$)	Holdout Error ($W \cdot kg^{-1}$)
CL1	2.84 [2.71, 2.97]	3.27	3.54
CL3	2.94 [2.86, 3.02]	3.24	3.56
CL4	3.26 [3.14, 3.38]	3.31	3.49
CL7	3.17 [2.87, 3.47]	3.25	3.55
CL8	3.47 [3.35, 3.59]	3.34	3.55
NC5	3.23 [3.14, 3.31]	3.25	3.60
NCA	3.24 [3.20, 3.28]	3.32	3.61

† Standard error range in brackets.
NCV = Nested Cross Validation; MCCV = Monte Carlo Cross Validation.

Table 5-11. Generalised predictive error for SVM models.

Cluster	NCV Error ($W \cdot kg^{-1}$) †	MCCV Error ($W \cdot kg^{-1}$)	Holdout Error ($W \cdot kg^{-1}$)
CL1	3.25 [3.14, 3.36]	3.24	3.46
CL6	3.33 [3.13, 3.53]	3.29	3.44
CL11	3.02 [2.87, 3.17]	3.27	3.45
CL7	3.43 [3.18, 3.68]	3.27	3.44
CL8	3.20 [3.03, 3.37]	3.25	3.47
NC5	3.18 [3.08, 3.28]	3.27	3.43
NCA	3.18 [3.13, 3.23]	3.29	3.44

† Standard error range in brackets.

Table 5-12. Generalised predictive error for GPR models.

Cluster	NCV Error ($W \cdot kg^{-1}$) †	MCCV Error ($W \cdot kg^{-1}$)	Holdout Error ($W \cdot kg^{-1}$)
CL1	2.92 [2.70, 3.14]	3.27	3.36
CL2	2.94 [2.83, 3.05]	3.24	3.32
CL3	3.00 [2.87, 3.13]	3.25	3.30
CL5	3.64 [3.51, 3.77]	3.38	3.48
CL6	3.25 [2.91, 3.59]	3.33	3.48
NC5	3.17 [3.06, 3.28]	3.66	2.87
NCA	3.17 [3.11, 3.23]	3.62	2.87

† Standard error range in brackets.

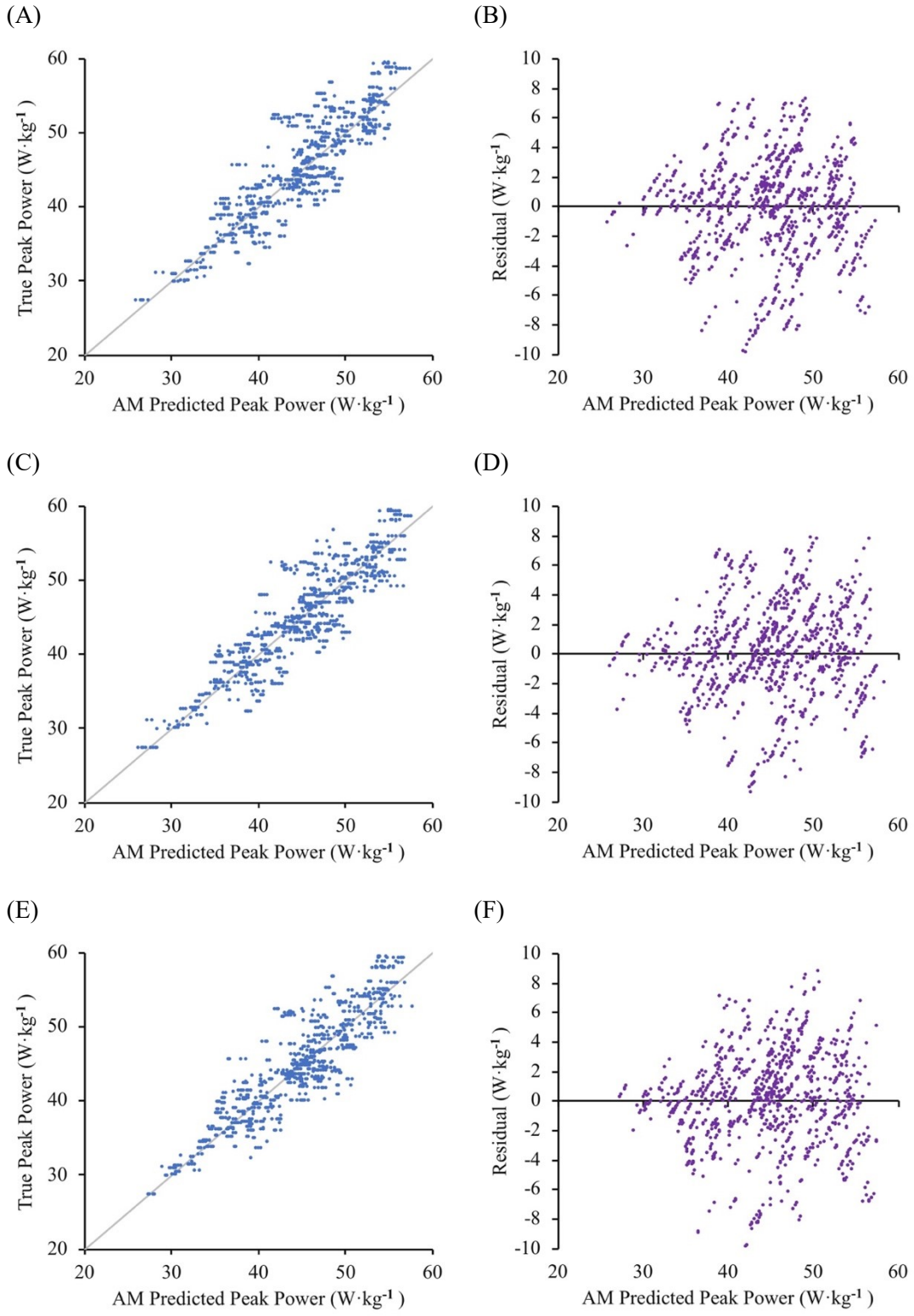


Figure 5-13. Model predictions for individual jumps for LR (CL3 model) (top row), SVM (CL1 model) (middle row) and GPR (CL2 model) (bottom row). Predictions according to the AM taken from 1000 × 10 MCCV, where for clarity 1000 random sampled points are shown from the 27552 generated. The apparent stratification where dots appear in series arises from the subsampling variation in the training set, subtly altering its prediction over multiple iterations for the same validation point.

5.4 Discussion

This chapter took the first step in developing an accelerometer model for predicting peak power in vertical jumping, building on the techniques developed and evaluated on gold-standard VGRF data in Chapter 4. The aim was to identify which of the data sets (sensor and jump type) were best suited for predicting peak power, to select suitable models (i.e. an algorithm with tuned hyperparameters), and to evaluate the models to test their predictive accuracy on unseen data. It involved developing a novel optimisation procedure within the framework of nested cross validation to comprehensively investigate the models in question. This was the first test of an FPCA-based accelerometer model to see how well it would perform with limited tuning before further refinements are made in the next two chapters.

5.4.1 Sensor attachment sites

It was established with multiple grid searches that models based on the LB sensor data consistently outperformed the equivalent models from other sensors. Indeed, the choice of attachment site accounted for more than half of the variance in model accuracy (53.3%). Many studies have used sensors attached to the lower back (e.g. Lesinski et al., 2016; Picerno, Camomilla, et al., 2011; Requena et al., 2012), although there is no consistency as the preferred location will depend on the application. It was supposed that attaching a sensor to the lower back would have the advantage of recording inertial accelerations closer to the body's CM. However, the CM does not have a fixed anatomical position but moves in space, and at various times may be outside the body. The action of the arms, when employed, will cause the CM to follow a more divergent path from a fixed anatomical location. The LB sensor's inertial accelerations will have reflected the flexion and extension at the ankle, knee and hip.

The models based on UB-sensor data were the second most accurate. This anatomical position is used by sensor systems favoured by many professional sports teams where players often wear an inertial sensor, which may also include GPS tracking. Located further from the fulcrum of the hip, the UB sensor would have experienced greater anteroposterior accelerations than the LB sensor in response to the trunk's changing inclination. The trunk initially tilts forward in the countermovement and then comes

upright rapidly during the propulsion phase. The UB sensor's inertial accelerations will have reflected not only the actions of multiple joints in the lower extremity, as with the LB sensor but also the actions of the erector spinae and other postural muscles. All these factors may have contributed to a more complicated acceleration pattern, which may explain why UB-sensor models were less accurate. The LB and UB models may have been more accurate than the models based on shank accelerations because the trunk makes the largest segmental contribution to the work done (Pandy et al., 1990). Luhtanen & Komi (1978) estimated that the trunk contributed 44% to the take-off velocity, while Miller & East (1976) concluded that the trunk contributed most of the total impulse. Hence, the inertial accelerations detected by the LB and UB sensors would be more correlated with the whole body's inertial movements.

In contrast, the shank-attached sensors were near the start of the kinetic chain, with only the ankle joint controlling anteroposterior as well as vertical accelerations. The joints of the foot would have played a minor role, with the further restriction of the training shoes inhibiting their actions. Hence, the shank acceleration patterns were simplified as there was no compound effect from the kinetic chain. The sensors would have experienced little soft-tissue movement underneath their position on the tibial aspect, unlike the LB and UB sensors that will have experienced considerably more extraneous movement. Attaching the sensors to the tibia allowed tape to be wrapped around the leg, exerting a strong compressive force to the sensor to minimise oscillatory movement. Wrapping tape around the waist for the LB sensor could not produce the same compressive force, such that some oscillation was possible in line with the spine. However, for the UB sensor located at C7, wrapping tape around the chest and shoulders was considered impractical. Consequently, the UB sensor will have experienced more extraneous movements.

In summary, each anatomical position had its advantages and disadvantages, but from the results, it seems clear that proximity to the body's CM was the critical factor. So whilst it might have been convenient to use the same UB-sensors worn routinely by professional rugby and football players, for example, the peak power estimates would be less accurate than those from LB sensors. Therefore, the following chapters will focus on the LB-sensor models to streamline the investigation. It is possible that the

subsequent developments may narrow the gap between the UB and LB models. However, the evidence from this chapter strongly points toward LB-sensor models retaining their advantage over their UB counterparts.

5.4.2 Jump type

The grid search also revealed that peak power in equivalent models for CMJ_A was harder to predict accurately than in the CMJ_{NA}, as expected. Arm swing influences the net forces acting on the body's CM, altering the CM's path throughout the jump. Mandic et al. (2015) showed that arm swing increases countermovement depth, which was also the case in the current study, as a further analysis confirmed using a mixed, random intercept model (0.023 m, $t = 12.7$, $p < 0.0001$). Moreover, Mandic et al. (2015) also demonstrated that peak power, unlike jump height, was sensitive to variations in countermovement depth. Samozino et al. (2008) improved the peak power prediction equation by introducing the squat depth as a third predictor (Section 2.3.2). This additional layer of complexity may explain why the CMJ_{NA} models were more accurate than those for the CMJ_A. That said, there were no great disparities between the algorithms in how their performance compared between jump types. These findings suggest that using FPCA to define the features can be effective when identifying characteristics related to the performance outcome in different jump types.

The grid search GLM established that jump type and sensor attachment site together explained the vast majority of the variance in model accuracy (87.1%), indicating modelling itself is less important than the practical considerations of the training environment and the needs of the sport. If the data set fails to capture the movement's essential characteristics, then there is little that model selection and hyperparameter tuning can achieve. That does not mean that expectations of further improvements should be tempered. Once the sensor and jump-type variation is removed, as later chapters will work with a single data set (LB-CMJ_{NA}), there will be considerably more apparent variation when manipulating other parameters. The larger point is that the fundamentals of data collection remain important, irrespective of the qualities of machine learning models. Care should be taken in choosing the sensor's anatomical

location, identifying it accurately on the body and attaching the sensor firmly and securely.

The analysis also showed that the choice of algorithm can be made independently of the jump type or the sensor attachment site, as the interactions of those two factors with the algorithm explained only 1.7% of the loss variance. Each condition had the same top three algorithms (SVM, LR and GPR) with similar relative differences between them. The algorithms may work in different ways, but they may be relying on similar features within the model, although perhaps with different emphasis. Therefore, it was concluded that model development should proceed by focussing on one data set (LB sensor data from the CMJ_{NA}). This will reduce the number of factors to consider, and hence the number of predictors in the model, with the strong likelihood that the conclusions drawn in the course of the investigation will apply to the other data sets, albeit with higher errors.

5.4.3 Model appraisal and comparison

Independent tests of the model's predictive ability came from the holdout estimate, a one-off test, and the more comprehensive NCV estimate. Summarised over all LR, SVM and GPR models, the NCV estimates of the validation RMSE were 2.84–3.64 $W \cdot kg^{-1}$ for all bagged models and 2.94–3.35 $W \cdot kg^{-1}$ for the largest clusters. When the models were retrained on the whole data set using the biggest clusters, the holdout validation error range was 3.32–3.56 $W \cdot kg^{-1}$. The non-nested MCCV estimate was the same for all three models at 3.24 $W \cdot kg^{-1}$. These results were based on an unbiased procedure, making comparisons with other studies less straightforward because they used non-nested cross validation or simply a holdout test. Given such considerations, the errors reported in this chapter still compare favourably to those for the peak power prediction equations (4.2 ± 2.1 $W \cdot kg^{-1}$, Figure 5-14, Section 2.3.2). As discussed in Chapter 2, this summary statistic was heavily influenced by one small investigation with only 20 participants, which reported a low error of 1.8 $W \cdot kg^{-1}$ (Canavan & Vescovi, 2004). It was therefore appropriate to devise a benchmark that took into account sample size (Section 2.3.2). That benchmark error of 4.6 $W \cdot kg^{-1}$ is somewhat larger than the generalised predictive errors reported in this chapter, ranging from

2.8 $W \cdot kg^{-1}$ to 3.6 $W \cdot kg^{-1}$. Moreover, those estimates were obtained with NCV, which yields an unbiased estimate of how the models could be expected to perform when applied to new data sets drawn from the same distribution. The holdout errors bore this out, ranging from 2.9 $W \cdot kg^{-1}$ to 3.6 $W \cdot kg^{-1}$. The NCV estimates were averages of multiple independent tests of the model, while the holdout estimate was from a single independent test. In contrast, when the peak power prediction equations were tested independently on different populations, the errors could be either very large or very small (Section 2.3.2): the Canavan-Vescovi equation was reported to produce errors of 27.6% (Lara et al., 2006), 25.3% (Amonette et al., 2012), or 2.0% (Lara-Sánchez et al., 2011); Harman's equation was reported to have errors of 18.5% (Lara et al., 2006), 20.9% (Amonette et al., 2012) or 3.8% (Quagliarella et al., 2010). The current study, on the other hand, produced more reliable error estimates with errors of 6.5%–7.2%. The NCV confidence intervals suggest that were independent researchers to follow the same procedures, there is a small chance one might obtain an RMSE of ~2 $W \cdot kg^{-1}$, while another one might find the error is ~4 $W \cdot kg^{-1}$. Even with this pessimistic forecast, the RMSE peak power is 9% which still compares favourably with many of the studies cited above.

It should be noted that the holdout tests were not fully independent in the sense that a different investigator did not gather the test data, and nor were the participants who were assigned to the holdout test set drawn from a separate group. Nevertheless, the approach combining nested cross validation with a holdout test was designed to evaluate the models rigorously and independently so far as such limitations allow. In fact, the participants who were assigned to the holdout set were chosen scrupulously at random using code run independently by another researcher. Therefore, these results and those in later chapters obtained with the same methods can be regarded with confidence. They provide the best possible estimate of how the models would perform on new data drawn from similar populations.

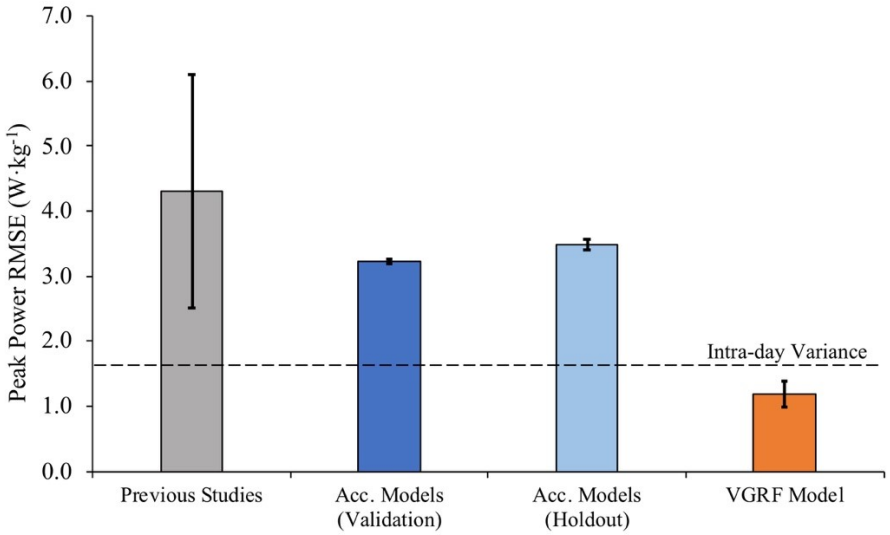


Figure 5-14. GPR models' peak power prediction errors in comparison with previous studies and the VGRF model from Chapter 4. The previous studies were (Ache-Dias et al., 2016; Amonette et al., 2012; Canavan & Vescovi, 2004; Lara et al., 2006; Quagliarella et al., 2010; Tessier et al., 2013). The intra-day variance is averaged from four reliability studies (Cormack, Newton, McGuigan, et al., 2008; Hori et al., 2007; McLellan et al., 2011; Taylor et al., 2012)

5.4.4 Review of the models

The parametric LR models assume a proportional, additive relationship between the predictors (FPC scores), among other strict assumptions. It is reasonable to suppose that the acceleration curve's amplitude may be proportional to peak power. A more rapid acceleration requires a greater force. When occurring late in the jump action, it would produce a higher instantaneous mechanical power. However, the implicit assumption is that the sensor's measurement of inertial acceleration is proportional to that of the body's CM, which is unlikely to be the case (Section 5.1). Nevertheless, the good performance of the LR model suggests that there must be some degree of correspondence. Thus, the additive nature of LR model would seem a reasonable choice for predicting peak power production from a biomechanical perspective. Regularisation would minimise the contribution of minor FPCs, as in the case of ridge regression or eliminate such FPCs in the case of the lasso method.

Taking the resultant, $|\mathbf{a}(t)|$, simplified the model, but in doing so, the model was unable to distinguish between different orthogonal directions. Although information

was lost, it may have been beneficial as the number of predictors were kept low (i.e. 15 FPCs rather than three times more). Nonetheless, taking account of the different orthogonal accelerations may help to improve predictions, even if there is no correction for the sensor changing orientation. The forces generated in the dynamic movement will not all be directed vertically, but will act to control and stabilise the body segments, particularly the trunk, which typically amounts to around half of the total body mass (Feltner et al., 2004; Nagano et al., 2007; van Soest et al., 1994; van Soest & Bobbert, 1993). Chapter 7 will introduce models based on the triaxial accelerometer data.

The preference for predictor standardisation to Z-scores in the SVM and GPR models indicated that the higher-order FPCs related in some useful way to the peak power output. However, those predictors only explained a comparatively small proportion of the acceleration curve variance. Still, the advantage of standardisation was relatively modest, as the parameter space plots revealed (Figure 5-12H, L), where the noise level from sub-sampling variation tended to have a greater effect, as represented by the surrogate model. This should not be surprising as the magnitudes of the FPCs diminish rapidly for higher orders. However, it should be remembered from Chapter 4 that VGRF FPC3 explained 72.9% of the variance in peak power despite it only accounting for 8.6% of the curve variance. Hence, it may be supposed that in respect of the accelerometer data, standardisation may help to bring to prominence higher order FPCs for the SVM and GPR models, which are heavily influenced by relative predictor magnitudes. The relative importance of the FPCs will be examined in Chapter 7, which considers feature selection.

The LR, SVM, and GPR models had predictive errors that were remarkably similar to one another, despite each model's internal workings being quite different. This finding suggests that the models identified similar information. Such similarity suggests that the models may be approaching the irreducible error, which depends on the data but not the algorithm (Rodríguez et al., 2010, 2013). This notion of an irreducible error, also known as the true error, suggests that the data preprocessing may not be optimal, placing a lower limit on the models' error. This being so, it follows that further improvements in accuracy may come from changes in data preprocessing. There will

be an upper limit to how much relevant information is available in the accelerometer signal, which, as noted above, depends on the anatomical attachment site and adherence to strict protocols.

5.4.5 Model cost, reliability and complexity

The computational cost of running the LR and GPR models were similar such that an optimisation involving 500 observations could take around half an hour, and with nesting, the full run could be expected to take 5–6 hours (iMac 2017, MacOS 10.14.6, 3 GHz Intel Core i5; Matlab R2020a ODE benchmark: 0.4620 s). However, for SVM, the fitting procedure could be lengthy, depending on the box constraint and kernel scale that could extend the localised region to include a large number of points for density estimation. It was not uncommon for a single m-th iteration (20 observations) to take around half an hour, and a single optimisation could run for half a day. The full run of 10 optimisations could take around five days. Such lengthy processing times are a serious drawback to using SVM, which did not yield any more accurate results than the other models. A nested cross validation that takes such a long time greatly impedes the research process as timescales are extended to weeks rather than days to learn if an approach has worked or not.

The SVM models occasionally produced an inordinately high loss that could have skewed the SM, destabilising the optimisation, were it not for a check to exclude such observations. This behaviour is arguably a severe limitation of the SVM model, given that more computational resources could be provided in principle to speed up the procedure. SVM was also a complex model as it depended on five hyperparameters, three of which had a strong effect on the loss (BC, KS and ϵ) (Figure 5-12). When the predictive accuracy of a complex model is similar to that of simpler models, it can be discarded in an appeal to Occam's razor and on the basis of the scientific method (Madigan & Raftery, 1994). The complexity argument is of practical concern because the number of parameters increases the dimensionality of the search space. Therefore, based on its unreliability, high cost and complexity, SVM models will be excluded from the subsequent investigations in this thesis.

5.4.6 Optimisation procedure

The novel optimisation procedure was introduced to overcome the difficulties of dealing with the variance in loss estimates from random subsampling variation (Hall & Robinson, 2009). It incorporated best practice techniques, including nested cross validation, random search, particle swarm optimisation and bagging. The optimisation procedure did not eliminate uncertainty, but it did provide a means by which equivalent models with similar predictive errors could be analysed and understood. In contrast, the Bayesian optimisation procedure could yield different optimal parameter values for the same data set (Appendix E.4). The Bayesian optimiser's parsimonious approach, which is a considerable advantage for expensive objective functions, can nonetheless result in a partial view of the parameter space. The new procedure was more successful because more observations from the random search helped smooth out the noise, reducing sensitivity to the peculiarities in the data (Rodríguez et al., 2010). It also gathered those observations without the overhead of an acquisition function that made the Bayesian optimiser prohibitively expensive (Appendix E.5). It retained the benefits of a Bayesian GP model while allowing a more comprehensive survey to be carried out. Since the SM optimum was the 'feasible minimum' rather than the lowest value observed, it provided an alternative to Tibshirani's standard error correction method (R. J. Tibshirani & Tibshirani, 2009).

The procedure was computationally intensive, but it is argued that having a survey of the parameter space helps investigators to understand the model rather than having to treat it is as a black box. An optimisation on its own may say a certain value is optimal, but it says nothing about whether other values may achieve similar levels of accuracy. That is why plots and surveys of the parameter space can be valuable as they can reveal sensitivity to certain parameters (Bischl et al., 2012; Cawley & Talbot, 2010). The benefits of a parameter space survey were limited in this chapter as it concerned abstract hyperparameters. However, this will become a powerful tool in the following chapters when optimising data preprocessing parameters that are more tangible.

5.4.7 Directed random search

The random search was made more efficient by directing observations towards regions of interest with lower loss estimates, a novel adaptation of the standard procedure. The approach proved highly effective as it brought down the number of observations required. Rather than the 5,000–10,000 observations needed using the standard algorithm only 500 were needed with the new method to obtain an optimal model. The improvement stemmed from the progressive reduction of the upper limit, $\mathcal{L}_{\max_m}$. This technique effectively directed the search, supplanting the role of the acquisition function from Bayesian optimisation. Blurring the upper limit allowed for the possibility of SM predictions being somewhat inaccurate, especially early on in the search, bearing in mind the SM was only updated every 20 observations. This flexibility allowed the search to effectively break out from the current region of interest, should the global minimum actually be located elsewhere. It effectively mimicked the ability of the Bayesian optimiser to avoid over-exploiting a region.

Specifying the number of observations required in advance set the schedule for the reduction in $\mathcal{L}_{\max_m}$. Too few observations would make the $\mathcal{L}_{\max_m}$ reduction too rapid, potentially directing the search to the wrong regions of the parameter space. For this chapter, 500 observations were found to be sufficient after some experimentation. In the next chapter, however, the number of parameters will be expanded. Additional observations may be required to be confident of finding the true optimum. Retaining weak parameters could also run the risk of overfitting the SM to nuances in AM performance (D. Baumann & Baumann, 2014), although bagging will help to mitigate this (Hall & Robinson, 2009). Therefore, in the next chapter, it would be sensible to exclude weak parameters from the optimisation.

5.4.8 Bagging

Bagging aggregated the multiple models produced in a given optimisation. Aggregating the model parameters and losses within and between outer folds yielded a more generalisable model. Fitting the SM's underlying GP was effectively another form of aggregation based on Bayesian principles. Choosing a subset of models to perform the aggregation was necessary at one level because the initial SM optima were

wayward and could not reasonably be considered for inclusion until the SM MAE dropped to a low stable value. It was then a question of whether to include all the models from that point onwards (NCA) to minimise standard errors or take the final five models of each series (NC5) on the assumption that the SM would have the highest fidelity with the AM in the final stages. A third possibility was to identify clusters of models (CL*n*) based on the evidence that certain distinct parameters values appeared to be favoured rather than there being a continuous distribution. The clustered models were better able to reflect nuance in the parameter landscape, where it existed (λ for LR, σ for GPR), and place the optima at the observable minimum in the SM plots (Figure 5-12C & K).

The clusters were defined solely on common parameters values without reference to the loss figures. Selecting a preferred clustered model should not be based on how well they perform in the outer fold for that would violate the principle of independence. Instead, the preferred cluster should be the largest one because that set of parameters was the one that the optimisation returned to the most, making the cluster model more robust to subsampling variations and hence more generalisable. It should be noted that the NC models did not go too far wrong, choosing a different value for the parameters that was still a reasonably good choice from the plots. λ is the best example where the loss was close to its lowest value over a wide range (Figure 5-12C).

5.4.9 Summary

The data set obtained from the LB sensor for the CMJ_{NA} gave rise to the most accurate models of peak power in vertical jumping compared to other combinations of anatomical location and jump type. This data set will be used in the following chapters as the accelerometer models are developed further. It was notable that the same algorithms performed well across data sets, providing reassurance that there should be no loss of generality by focusing on the LB-CMJ_{NA} data set alone. Hence, the findings of subsequent investigations should be applicable to the other data sets, including those from the UB sensor that is popular in professional and elite sport.

The optimisation procedure showed that the LR, SVM and GPR models produce peak power estimates with very similar levels of accuracy. Based on nested cross validation,

the predictive errors of 2.8–3.6 $W \cdot kg^{-1}$ were equivalent to 6.2-8.0% based on the mean peak power. Independent tests based on the holdout data set reported slightly higher errors of 3.3–3.6 $W \cdot kg^{-1}$. These FPCA-based accelerometer models were more accurate than most but not all of the previously published peak power prediction equations (4.3 ± 1.8 $W \cdot kg^{-1}$). Only the LR and GPR models will be taken forward to the next chapter. SVM was less reliable, more complex and computationally expensive, which could not be justified given its results were no better than the other two.

The novel optimisation procedure introduced in this chapter was thorough and comprehensive, bringing together several well-established procedures. The surrogate model had the same design as Matlab's Bayesian optimiser. It was more robust and reliable because it was trained on many more observations. Crucially, the same random search that gathered the points to map the parameter space was used for the optimisation, bringing consistency to the results. It will serve as the foundational method for the research in the rest of this thesis. The full benefits of this approach will become more apparent in the next chapter, where the optimisation procedure will be applied to the parameters governing data preprocessing.

CHAPTER 6. DATA PREPROCESSING

6.1 Introduction

The development of an accelerometer model began in the previous chapter by identifying the jump type, sensor attachment site and algorithms that were most conducive to making accurate peak power predictions. The shortlisted models had a similar predictive error suggesting that they may have been limited by the information available. This chapter investigates how further information can be extracted from the accelerometer data to reduce the predictive errors. Specifically, it addresses the third research question, *how should the accelerometer data be preprocessed to minimise the model's predictive error?*

The accelerometer data was truncated at take-off in the previous chapter, reflecting the calculations of jump height and peak power, but there is no reason why that period should apply to the accelerometer data if extending it provides more relevant information. Studies into sports-related applications of wearable sensors have often chosen time windows that extend either side of the event of interest, typically employing time windows of ±1 s, ±2 s or sometimes up to ±5 s (Blank et al., 2015; Groh et al., 2017, 2016, 2016; Kautz et al., 2017; McGrath et al., 2019; Rawashdeh et al., 2015). In vertical jumping, take-off would usually be thought of as the event of interest, but extending the time window raises the possibility of realigning the curves at some other point. The jump landing may be a suitable alternative or the timing of the acceleration impact spike, which may be identified more easily. Algorithms to locate such events should be based on the accelerometer signal alone so the system envisaged for jump testing can be self-reliant. In this way, it can be low-cost and suitable for field testing.

Smoothing the accelerometer signals can remove or retain information depending on the roughness penalty employed (Levitin et al., 2007; Ramsay & Silverman, 2005). Adjusting the roughness penalty is a form of regularisation, controlling the degrees of freedom inherent in the time series. It is analogous to low-pass signal filtering that is often applied in biomechanics research, including machine learning applications (e.g.

Blank et al., 2015; Groh et al., 2017; McGrath et al., 2019; Mlakar and Luštrek, 2017; Zago et al., 2019). Regularisation applied to the data implicitly controls the degree of complexity in the subsequent model once the features have been extracted. Hence, it applies to all model types, although each algorithm has some form of regularisation built-in (λ for LR, ε for SVM and σ for GPR). The number of FPCs is another factor in controlling model complexity that will be addressed in the next chapter. The main consideration for the investigation below is optimising the parameters governing the time window and the level of smoothing.

The final preprocessing step before running FPCA is curve registration. If suitable landmarks can be identified, the acceleration curves can be more closely aligned, allowing FPCA to extract more meaningful information. When registration was evaluated on the VGRF data, the landmarks identifying the zero power and peak power were the most successful (Table 4-2). However, registration was only beneficial when the curves were out of alignment to begin with, as was the case with time normalisation. Although the accelerometer data is padded, it is more variable than the VGRF data such that further steps to alignment the signals' peaks and troughs may be necessary, and hence beneficial to the model. Therefore, curve registration should be investigated to determine it enhances the peak power models based on accelerometer data.

In summary, this chapter aims to:

- Determine the optimal settings for preprocessing the accelerometer data in terms of the time window, the level of smoothing, the appropriate alignment point, and whether and how to perform curve registration.

In so doing, it will be necessary to develop algorithms for identifying the jump take-off, landing and impact points in the accelerometer signals to permit self-sufficiency for an accelerometer-based system. In light of the results produced, algorithm selection can be revisited.

6.2 Methods

The AM function incorporated all data preprocessing operations so the optimiser could specify changes, as required, to the time window, the alignment point and the level of smoothing. Fulfilling the aims of this chapter was therefore a matter of defining the new data preprocessing parameters and setting up the optimisations appropriately. The methods described in Chapter 5 were followed, with the only new developments being algorithms to identify take-off, landing and impact for signal alignment and a further procedure to locate the landmarks for curve registration. An initial optimisation identified non-significant parameters that were removed in a second optimisation. The following sub-sections describe these elements in more detail.

6.2.1 Event-detection algorithms

The event-detection algorithms were based on the raw accelerometer signals. The criterion times for take-off and landing were when the VGRF first fell below 10 N and then when it first rose above this threshold. The impact point had no criterion value as such because it was obtained directly from the accelerometer data.

The landing-detection algorithm was executed first because the impact acceleration spike was easily identifiable using the resultant signal. With reference to Figure 6-1, the algorithm:

1. Smoothed the raw data using a centred, 11-point moving average;
2. Identified candidates for the peak associated with the landing impact from all maxima above 1 g (‘*Freefall Threshold*’);
3. Excluded peaks not preceded by a quiescent period where the signal fell below 1 g in 120 ms, indicative of freefall (points 1B and 1C);
4. Recorded the point immediately before each of the shortlisted candidate peak(s) where the signal rose above 1 g (‘*Impact Threshold’*) (point 1A)
5. Calculated the mean acceleration over 210 ms before each of these points, called the ‘*Freefall Mean*’ (see the shared region);

6. Selected the peak with the highest ratio of peak value to *Freefall Mean*, designated as the *Impact Point* (which turned out to be 1A);
7. Placed the *Landing Point* 40 ms before the *Impact Threshold* (point 2). A fixed offset from the *Freefall Threshold* proved to be more accurate than the last minimum in the freefall region.

These steps were devised from an inspection of the curves. The numbers quoted above were determined using the Matlab Bayesian optimiser, which minimised the RMSE between the estimated landing time and the criterion landing time. The optimiser's objective function returned the mean error of ten bootstrapped iterations of the algorithm in order to obtain a more generalisable algorithm, not one simply optimised for this particular dataset.

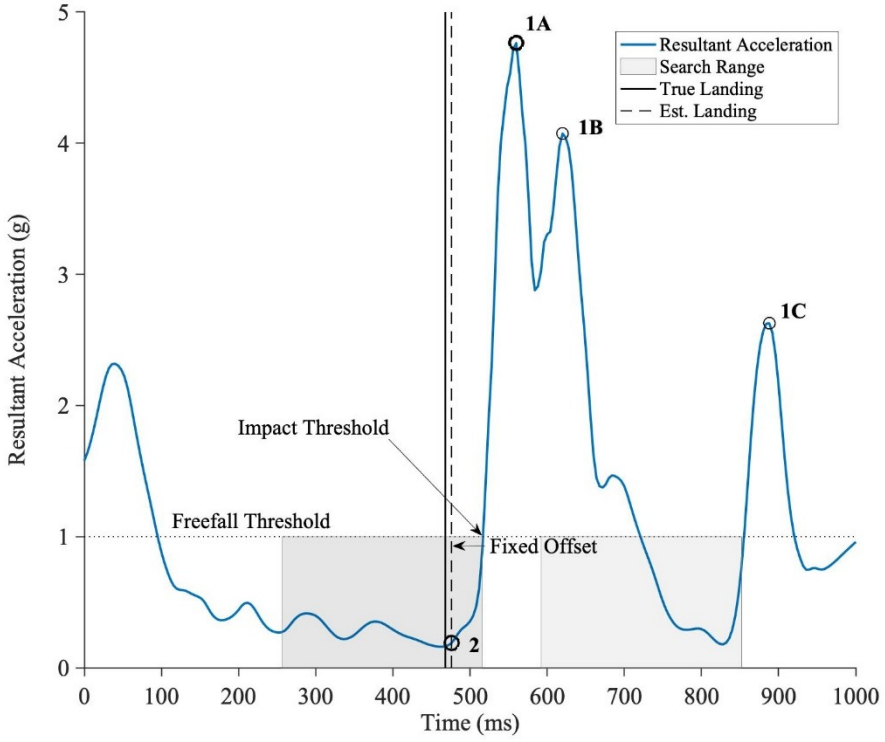


Figure 6-1. Landing algorithm illustrated with an example of the resultant acceleration signal from the LB sensor. Time is measured from the true take-off in milliseconds. The candidate maxima for the impact peak are indicated by 1A, 1B and 1C. The Freefall Mean is the mean acceleration over a shaded region. The best candidate (1A) is preceded by a low acceleration, from which the landing estimate is made (2).

The take-off detection algorithm relied on the X and Z axes of the orthogonal acceleration signals. The signals were terminated at the previously determined landing point to avoid false detections in the landing part of the signal. With reference to Figure 6-2, the algorithm:

1. Rotated the signal so that the inertial acceleration was approximately vertical at the start (-1 g, 0, 0) when the participant was quietly standing still;
2. Smoothed the X and Z acceleration time series with a central, 21-point moving average;
3. Differentiated the smoothed signals using a 2-point central difference formula to produce the jerk time series, called *DX* and *DZ*, respectively;
4. Found the first X maximum after the first X minimum (‘*XPeak’*) (points 1 and 2);
5. Located the nearest DX and DZ peaks before XPeak (‘*DXPeak’* and ‘*DZPeak’*, respectively – (points 3 and 4)
6. Select *DZPeak* as the take-off point if *DZPeak* and *XPeak* were separated by less than 340 ms; otherwise, it chose *DXPeak* after adding a 4 ms offset.

As with the landing-detection algorithm, the steps were devised from inspection, but the figures quoted above were determined using the Bayesopt optimiser. The signal rotation (16.4 ± 7.8°, mean ± SD) improved the algorithm’s accuracy, which relied on the X and Z accelerations at take-off being aligned to the vertical and posterior direction, respectively. *XPeak* could not be identified consistently without the data truncated at landing. Take-off was more likely to coincide with the peak jerk in the posterior (DZ) rather than the vertical direction (DX).

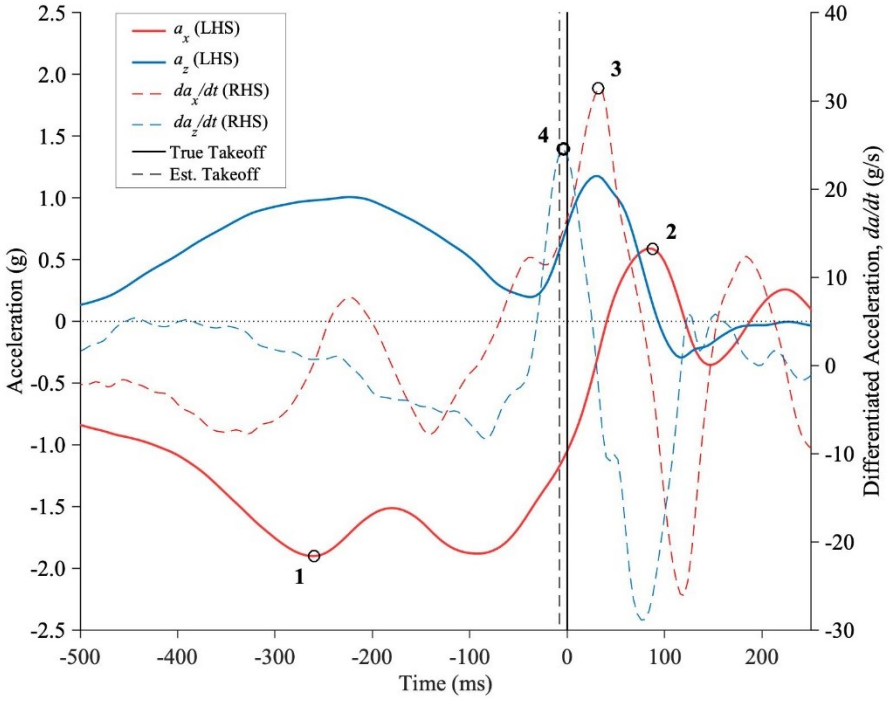


Figure 6-2. Take-off algorithm illustrated with an example of the X and Z acceleration time series, including the differentiated DX and DZ curves. The algorithm progresses in sequence identifying the X minimum (point 1), X maximum (point 2), and the DX and DZ maxima (points 3 and 4). DZ is selected as the reference point for take-off. The timescale terminates at landing, determined by the Landing algorithm.

The error from using the take-off detection algorithm instead of the criterion VGRF times was assessed using the Chapter 5 models (CL3 cluster model for LR and CL2 for GPR). First, a simple comparison was made between the two methods using 1000 × 2 MCCV to determine the change in the validation error. Second, the models' sensitivity to signal misalignment was determined to discover the benefit that might be gained with a better event-detection algorithm. Gaussian noise was introduced, centred on the VGRF-determined take-off points (Section 3.2.5). The standard deviation was increased from 0 ms to 100 ms in steps of 4 ms, the smallest increment possible with a 250 Hz sampling frequency. Random noise was re-generated 100 times for each step. The LR and GPR models' validation errors were averaged over 100 × 2 MCCV, making 10,000 observations in total for each model.

6.2.2 Model parameters

The AM function was extended to include parameters specifying the alignment event, the time window and the functional smoothing (Table 6-1). As in Chapter 5, the resultant accelerometer signal was used with 15 FPCs retained. The time window was defined by t_{pre}, the period before the alignment event, and t_{post}, the corresponding period after it. Together these parameters, in conjunction with the alignment event, defined the start and end of the time series. Since the time was discretised, t_{pre} and t_{post} were treated as integers in the optimisation so that each increment represented 4 ms. For example, in Chapter 5, $t_{\text{pre}} = 500$ intervals (2000 ms) and $t_{\text{post}} = 0$, where the alignment event was take-off. In Table 6-1, the lower bounds of t_{pre} and t_{post} are negative to prevent boundary effects from the SM's Gaussian Process (arising from the assumption of a constant basis function) influencing the optimisation.

Table 6-1. Optimisation parameters for LR and GPR, including the new parameters introduced in Chapter 6.

Model	Parameter	Type †	Hyperparameter Values	PSO Bounds ‡
LR	Regularisation	C	{Ridge, Lasso}	[0.50, 2.49]
	Solver	C	{SVM, Least Squares}	[0.50, 2.49]
	Regulariser, λ	R	10^{-12} … 10^{12}	[-12, 12]
GPR	Basis	C	{None, Constant, Linear, Pure Quadratic}	[0.50, 4.49]
	Kernel	C	{Squared Exponential, Exponential, Matérn 3/2, Matérn 5/2, Rational Quadratic}	[0.50, 5.49]
	Noise, σ	R	10^{-4} … 10^{2}	[-4, 2]
	Standardisation §	C	{No, Yes}	[0.50, 2.49]
Both	Alignment	C	{Takeoff, Landing, Impact}	[0.50, 3.49]
	t_{pre} *	I	{-50, …, 750}	[1, 801]
	t_{post} *	I	{-50, …, 750}	[1, 801]
	No. Basis Functions	I	{15, …, 200}	[15, 200]
	Roughness penalty, λ_{FS}	R	10^{-12} … 10^{12}	[-12, 12]

† Parameter type: C = Categorical (nominal variable); R = Real (continuous), I = Integer (discrete).
‡ Categorical parameters are indexed; real parameters are log-transformed.
§ Standardise the predictors and outcome variables as Z scores during fitting.
* Time intervals were in 4 ms units giving a range -200 ms to 3000 ms.

The smoothing parameters specified the number of b-spline basis functions and the roughness penalty, λ_{FS}. In the last chapter, 100 basis functions were used with $\lambda_{FS} = 10^2$. λ_{FS} was allowed to vary over 25 orders of magnitude to cover a wide range of possible levels of smoothing. The number of basis functions ranged from a minimum of 15 basis functions (determined by the number of FPCs retained in FPCA) to an upper limit of 200 that capped the computational costs. This cap could be raised in the second optimisation if the results justified it.

6.2.3 Statistical analysis and optimisation

It was noted in Section 5.4.7 that some parameters could substantially affect the predictive error (e.g. λ in LR) whilst others may only have a weak effect (e.g. regularisation method in LR). Including weak parameters in an optimisation would run the risk of overfitting if those variables were largely uncorrelated with the model's predictive error (D. Baumann & Baumann, 2014). Overfitting was also a stronger possibility in this chapter because the optimisation involved more parameters (Cawley & Talbot, 2010). Hence, it was essential to contain the complexity as far as possible by eliminating weak parameters from the optimisation (Figure 6-3).

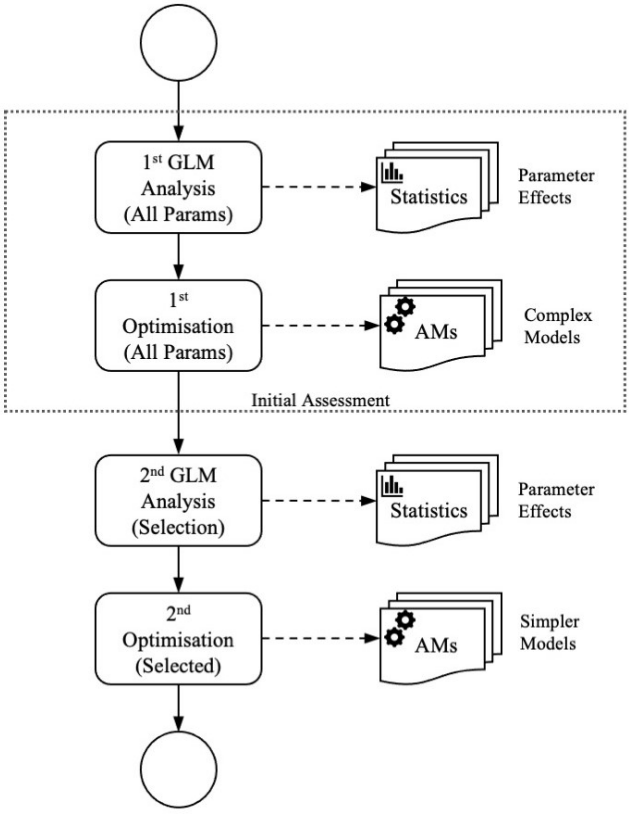


Figure 6-3. Overview of the two-stage process followed in this chapter to simplify the models. At both stages, a statistical analysis is performed to assess each parameter's effect on the models' predictive error.

A separate random search was conducted for each analysis to gather 100 observations on the AM loss for each of the 10 outer folds (1000 in total), drawn from the parameter ranges specified in Table 6-1. The AM loss was estimated using the same inner cross validation design, as before: 20 × 2 MCCV. The points were chosen purely at random to make them independent, unlike in an optimisation where they would be concentrated progressively. A statistical analysis was performed using a GLM with the AM parameters as the predictor variables and the AM inner loss as the outcome variable. The inner loss was preferable to the outer loss because it had a direct relationship with the parameters. The coefficients of the linear GLM indicated the overall effect of each parameter, which for categorical parameters was broken down by level. The total variation in the AM loss explained by each parameter was estimated using the semi-partial ω^2.

The first-round GLM analysis was conducted with all parameters to quantify their effects. Then an optimisation was carried with all parameters following the nested cross validation methods described in Sections 5.2.5–5.2.10. The number of random search observations, $\mathbf{Z}^{(k)}$, was doubled to $I = 1000$ for each outer fold to accommodate the higher number of dimensions of the parameter space. Accordingly, each $\mathrm{SM}_m^{(k)}$ series comprised $M = 50$ models.

A second-round GLM analysis was then performed using the same stepwise selection method to select the VGRF FPCs (Section 4.2.5). The random search, which provided the data set for the GLM, was re-run with all parameters except alignment because take-off was the overwhelming choice from the first optimisation (Figure 6-5). The GLM-selected parameters were then put into a second-round optimisation with the same setup as the first.

6.2.4 Curve registration

Once the optimal parameter values were known, the final stage was to introduce curve registration to determine whether it could improve the models' accuracy. It was computationally prohibitive to perform curve registration within the optimisation because the operation itself was very intensive, requiring approximately five minutes of processing time for every AM function evaluation. Curve registrations were therefore computed in advance based on a grid of predetermined functional smoothing parameter values. Since the number of basis functions proved to be a weak parameter, it was held constant at 100 basis functions, and only the roughness penalty was discretised with intervals of $\log_{10} \Delta\lambda_{\mathrm{FS}} = 0.2$ over the range $\log_{10} \lambda_{\mathrm{FS}} = \{0, \ldots, 10\}$. Since the landmarks were the cross-sectional means, one registration could be used across all data subsamples.

There were no suitable landmarks in the acceleration curves themselves, but the derived pseudo power curves did have prominent peaks around take-off and landing, which were located as follows. The resultant acceleration signal was converted into a smooth function comprising 50 fourth-order, b-spline basis functions with a roughness penalty of $\lambda_{\mathrm{FS}} = 10^4$. The smoothing parameters were fixed and independent of the corresponding parameters set during the optimisation. The algorithm then converted

the functional curves back into a discrete set of points, sampled at 1000 Hz. It then integrated the curve using the cumulative trapezoidal rule to obtain the velocity, assuming a zero initial velocity since the participant stood still before jumping. It then multiplied the corresponding points of the (smoothed) acceleration and velocity time series to produce the power time series, where the acceleration serves as a proxy for force. The inverse proportionality with body mass was ignored as the correct scale was irrelevant when finding peaks. The two peak power landmarks (L_1 and L_2) were on either side of the landing timing point found earlier (Section 6.2.1). L_1 was not always the peak immediately before the landing, so it was defined as the peak with the highest ratio of prominence to proximity, based on Matlab's **findpeaks** function. L_2 was simply the first maximum after landing.

The landmark set parameter was introduced to the optimisation specifying the landmarks to use, if any: {None, L_1, L_2, L_1 & L_2}. The best performing model from the second round optimisation was chosen to evaluate registration. t_{pre} and t_{post} were held constant at the times determined in the second optimisation. λ_{FS} was treated as a discrete parameter rather than a continuous one with the range and interval spacing that matched registration preprocessing. The other remaining parameters were allowed to vary as before.

6.2.5 Flight time

The influence of flight time on the models was investigated once the optimisation was complete, given that flight time had been a predictor in the peak power prediction equations (Section 2.3.2). If the optimal time window encompassed take-off and landing, an analysis would be needed to determine whether the models relied on the flight time or the landing acceleration patterns. It was based on comparing how the model performed with flight time allowed to vary freely (no intervention) with an artificial situation in which flight time was held constant by padding out the flight phase. The fixed flight time was set at 1000 ms to ensure flight times from all jumps were extended.

The acceleration time series were padded with the resultant signal's minimum value after take-off starting at the point where it was achieved. Thus, the curve's lowest point

was extended by the required number of points, making the minimum appear as a flat section of the curve. In choosing the lowest point, the intervention had the smallest possible influence on the acceleration curves. The minimum was located using an 11-point moving average to ensure the algorithm reliably identified the point when the body was closest to being in freefall, according to the accelerometer.

The effect on the LR and GPR models' predictive error in both cases (actual and fixed flight times) was investigated by using a grid search to increment t_{post} in units of 4 ms from zero up to 2000 ms while t_{pre} was held constant at 2000 ms. If the validation error dropped when the time window encompassed landing in both cases, then the models would be using the landing acceleration patterns. If the error only dropped in the non-intervention case (actual landing times), then the model was reliant on the flight time.

6.3 Results

6.3.1 Performance of jump detection algorithms

The algorithms to detect jump take-off and landing had RMSEs of 11.9 ms and 11.7 ms (~ 3 time intervals each), respectively, compared to the criterion timing of those events based on the VGRF data (Section 3.2.5). These errors amounted to a degree of misalignment between the accelerometer signals, which increased the Chapter 5 models' validation errors by a 0.09 $W \cdot kg^{-1}$ and 0.15 $W \cdot kg^{-1}$, respectively for LR and GPR (Table 6-2). There was no error as such for the acceleration impact spike because it was obtained directly from the signal.

Table 6-2. Impact of switching to identifying take-off based on the accelerometer data rather than having to rely on the criterion VGRF data. All figures in $W \cdot kg^{-1}$.

Model	VGRF-based Alignment	Accelerometer-based Alignment	Difference
LR	3.48	3.57	+0.09
GPR	3.49	3.64	+0.15

† Chapter 5 models (largest cluster) – LR: CL3; GPR: CL2.

The models' sensitivity to temporal misalignment based on Gaussian noise only became apparent when the SD ≥ 12 ms (Figure 6-4). The LR model proved to be more sensitive than GPR to larger misalignment errors. The simulated misalignment under-estimated the actual error of both models (disparities of 0.04 $W \cdot kg^{-1}$ and 0.10 $W \cdot kg^{-1}$), but those estimates were within the SD range in Figure 6-4. The consequent effect on model predictive error was considered minor (0.3%) in the context of the broader results presented in this thesis. It was concluded that the take-off detection algorithm could be accepted and incorporated as a permanent feature of the AM function. The landing and impact-detection algorithms were expected to have a similar influence on model predictive error since their misalignment errors were almost identical, an assumption that was tested in the subsequent optimisations.

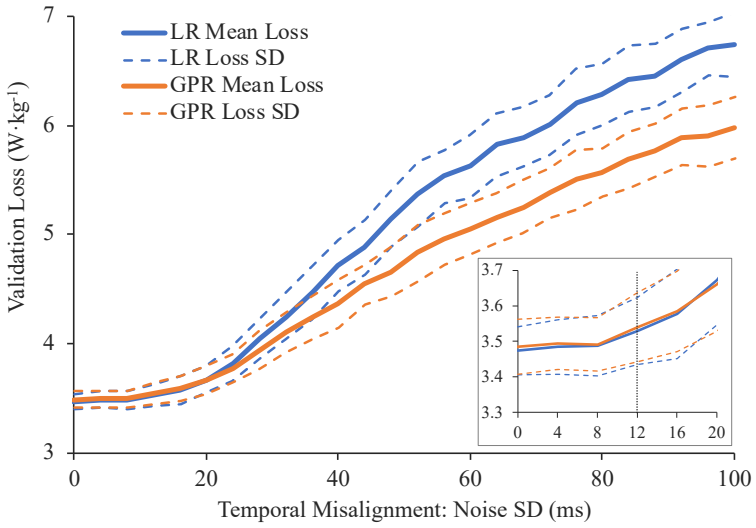


Figure 6-4. Impact on validation loss of temporal misalignment of the accelerometer signal with the true take-off time determined from the VGRF data. The misalignment was simulated by Gaussian noise, defined by its SD. Inset: narrow-range scale showing the algorithm's actual misalignment error indicated by the vertical dotted line at 12 ms.

6.3.2 GLM statistical analysis

The GLM based on all LR model parameters explained 61.2% of the variance in AM loss. In comparison, the second GLM based on only the selected parameters (omitting alignment) explained 63.3% of the variance (Table 6-3). The LR regularisation parameter, λ, accounted for almost all of this variance in both cases (55.9% and 61.7%, respectively). The three other selected parameters were the time window parameters, t_{pre} and t_{post}, and the roughness penalty, λ_{FS}. In the first optimisation, t_{pre} appeared to be the more influential time parameter, but the second GLM showed that t_{post} was, in fact, more important when the potential confounding effects of weak parameters were eliminated. λ_{FS} was significant, but the number of basis functions had almost no effect. With signal alignment at take-off, the LR model estimates were 0.26 $W \cdot kg^{-1}$ lower than if alignment had been at the landing or impact points. The ridge regularisation remained the preferred method, as in Chapter 5, but least squares was now the better solver.

Table 6-3. GLM analysis of LR model for the first and second optimisations, where the latter includes only selected parameters. The effect on the AM validation loss is shown, together with the *F*-statistic (ratio of explained to unexplained variance) and the proportion of total variance explained.

GLM Model		1st Optimisation			2nd Optimisation		
LR Parameter	**Levels**	**Effect (W·kg^{-1}) †**	***F***	**ω^2 ‡**	**Effect (W·kg^{-1}) †**	***F***	**ω^2 †**
Intercept		6.256			6.234		
Regularisation	Lasso	+0.152	1.1	0.0%			
	Ridge	.					
Solver	SVM	.	4.5	0.1%			
	Least Squares	-0.123					
$\log_{10}\ \lambda$		+0.185	1440 ***	55.9%	+0.195	1700 ***	61.7%
Alignment	Take-off	.	9.8 ***	0.7%			
	Landing	+0.266					
	Impact	+0.265					
t_{pre} (s^{-1})		-0.349	97.6 ***	3.8%	-0.100	8.9 *	0.3%
t_{post} (s^{-1})		-0.052	3.1	0.1%	-0.200	32.2 ***	1.1%
No. Basis Functions		0.000	0.4	0.0%			
$\log_{10}\ \lambda_{FS}$		+0.019	16.8 ***	0.6%	+0.019	15.1 **	0.5%
TOTAL				**61.2%**			**63.3%**

GLM RMSE = 1.03 W·kg^{-1} (1st optimisation); 1.05 W·kg^{-1} (2nd optimisation).

† GLM coefficient. Intercept is the overall mean loss. Coefficients make adjustments per unit value: e.g. for logarithmic parameters, per unit change in the exponent, for time parameters, per second. For categorical parameters, relative adjustment shown for different levels. Reference point for these adjustments ('.' gaps) is the Chapter 5 model.

‡ Semi-partial ω^2, the proportion of the total variance explained.

* $p < 0.01$; ** $p < 0.001$; *** $p < 0.0001$. No asterisk indicates non-significant: $p > 0.01$.

Type-I estimates quoted. *F* statistic relates to semi-partial ω^2.

The GLM for the GPR model explained a much smaller proportion of the AM loss variance, 25.4% with all parameters included and 21.3% based on the selected parameters (Table 6-4). The selected parameters, those with a strong influence on model prediction, were the basis function, σ, t_{pre}, t_{post} and λ_{FS}. In the second GLM, t_{post} explained the highest proportion of the variance (7.4%), but this was modest compared to the dominant role played by λ in the LR model. The basis function was the second most important parameter, where the linear function now appeared to be most advantageous, compared to no basis function in Chapter 5. The GLM coefficients were larger for the time window parameters than the corresponding coefficients for the LR model, suggesting that the GPR model was more sensitive to the chosen demarcation points.

Table 6-4. GLM analysis of GPR model for the first and second optimisations, where the latter includes only selected parameters. The effect on the AM validation loss is shown, together with the *F*-statistic (ratio of explained to unexplained variance) and the proportion of total variance explained.

GLM Model		1st Optimisation			2nd Optimisation		
LR Parameter	**Levels**	**Effect (W·kg^{-1}) †**	***F***	**ω^2 ‡**	**Effect (W·kg^{-1}) †**	***F***	**ω^2 †**
Intercept		5.938			5.644		
Basis Function	None	.	25.0 ***	5.4%	.	31.3 ***	6.3%
	Constant	-0.243			+0.236		
	Linear	-0.997			-0.571		
	Pure Quadratic	-0.273			+0.455		
Kernel Function	Squared Exp.	+0.488	6.2 ***	1.5%			
	Exponential	+0.129					
	Matérn 3/2	-0.005					
	Matérn 5/2	+0.256					
	Rational Quadratic	.					
$\log_{10}\ \sigma$		+0.152	48.0 ***	3.5%	+0.166	45.9 ***	3.5%
Standardisation	No	+0.054	0.5	0.0%			
	Yes	.					
Alignment	Take-off	.	11.3 ***	1.5%			
	Landing	+0.436					
	Impact	+0.461					
t_{pre} (s^{-1})		-0.525	150 **	11.1%	-0.175	13.1 **	1.0%
t_{post} (s^{-1})		-0.145	11.7	0.8%	-0.450	95.4 ***	7.4%
No. Basis Functions		0.000	0.0	-0.1%			
$\log_{10}\ \lambda_{FS}$		+0.027	22.6 ***	1.6%	+0.037	40.9 ***	3.1%
TOTAL				**25.4%**			**21.3%**

GLM RMSE = 1.24 W·kg^{-1} (1st optimisation); 1.27 W·kg^{-1} (2nd optimisation)

† GLM coefficient. Intercept is the overall mean loss. Coefficients make adjustments per unit value: e.g. for logarithmic parameters, per unit change in the exponent, for time parameters, per second. For categorical parameters, relative adjustment shown for different levels. Reference point for these adjustments ('.' gaps) is the Chapter 5 model.

‡ Semi-partial ω^2, the proportion of the total variance explained.

* $p < 0.01$; ** $p < 0.001$; *** $p < 0.0001$. No asterisk indicates non-significant: $p > 0.01$. Type-I estimates quoted. *F* statistic relates to semi-partial ω^2.

6.3.3 Optimisation performance

The second round of optimisations using the selected parameters identified above produced optimal models that were generally more accurate than in the first round (Table 6-5). The GPR models at both stages outperformed the LR models in all metrics presented below. It was also evident that the best clustered models outperformed the more general bagged (non-clustered) models. The similarity between NCV and MCCV estimates provided support for the novel optimisation method. Signal alignment at take-off was the overwhelming choice for both models (Figure 6-5).

Table 6-5. Summary results of the first and second round optimisations for LR and GPR models, showing representative loss estimates by different metrics.

	LR Optimisation		GPR Optimisation	
Loss Estimate ($W \cdot kg^{-1}$)	**1st Round**	**2nd Round**	**1st Round**	**2nd Round**
NCV (Bagged)	3.31	3.20	3.08	2.75
NCV (Best Cluster)	2.81	2.97	2.69	2.47
MCCV (Bagged)	3.36	2.97	2.82	2.77
MCCV (Best Cluster)	2.97	2.96	2.74	2.64

NCV = Nested Cross Validation; MCCV = Monte Carlo Cross Validation.

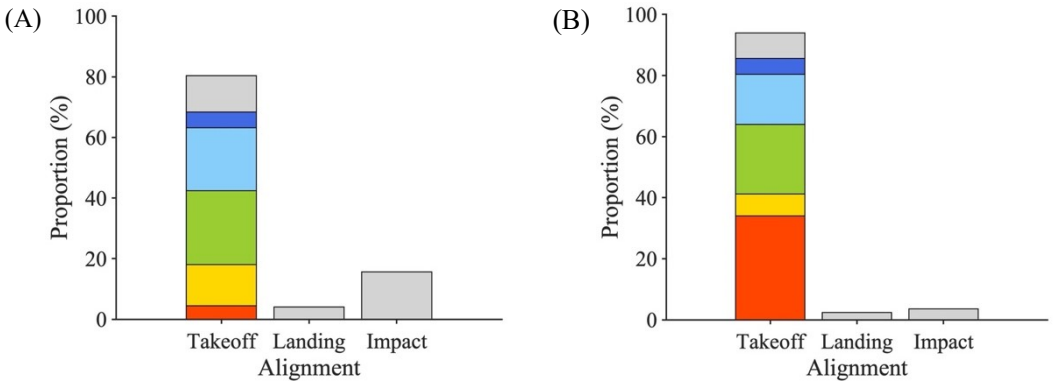


Figure 6-5. First round optimisation selection of accelerometer signal alignment for LR (A) and GPR (B) models. Colours indicate clusters.

The second round optimisations converged in fewer iterations than the first owing to the search space having fewer dimensions. The surrogate models had a higher fidelity with the AM, as indicated by a lower MAE (0.14 $W \cdot kg^{-1}$for LR and 0.13 $W \cdot kg^{-1}$ for GPR), a smaller confidence interval (0.34 $W \cdot kg^{-1}$ and 0.32 $W \cdot kg^{-1}$, respectively) and a narrower gap to the noise level (0.03 $W \cdot kg^{-1}$ in both cases). The surrogate model MAE had stabilised after 420 and 480 search observations, respectively, setting the start point for cluster analysis and bagging.

The spread of AM loss estimates for individual clusters is presented in Figure 6-6. The inner loss had a small variance compared to the outer loss for both LR and GPR models, indicating that the model's performance can be much more variable when subject to an independent test. The outer loss variance was slightly narrower for the second optimisation, particularly for the GPR model. The outer AM distributions were skewed in some cases with the mean indicator and the median line drawn some way from the middle of the box. The GPR models, CL3 and CL4, stand out for their narrow spreads but suffer from several outliers (Figure 6-6H).

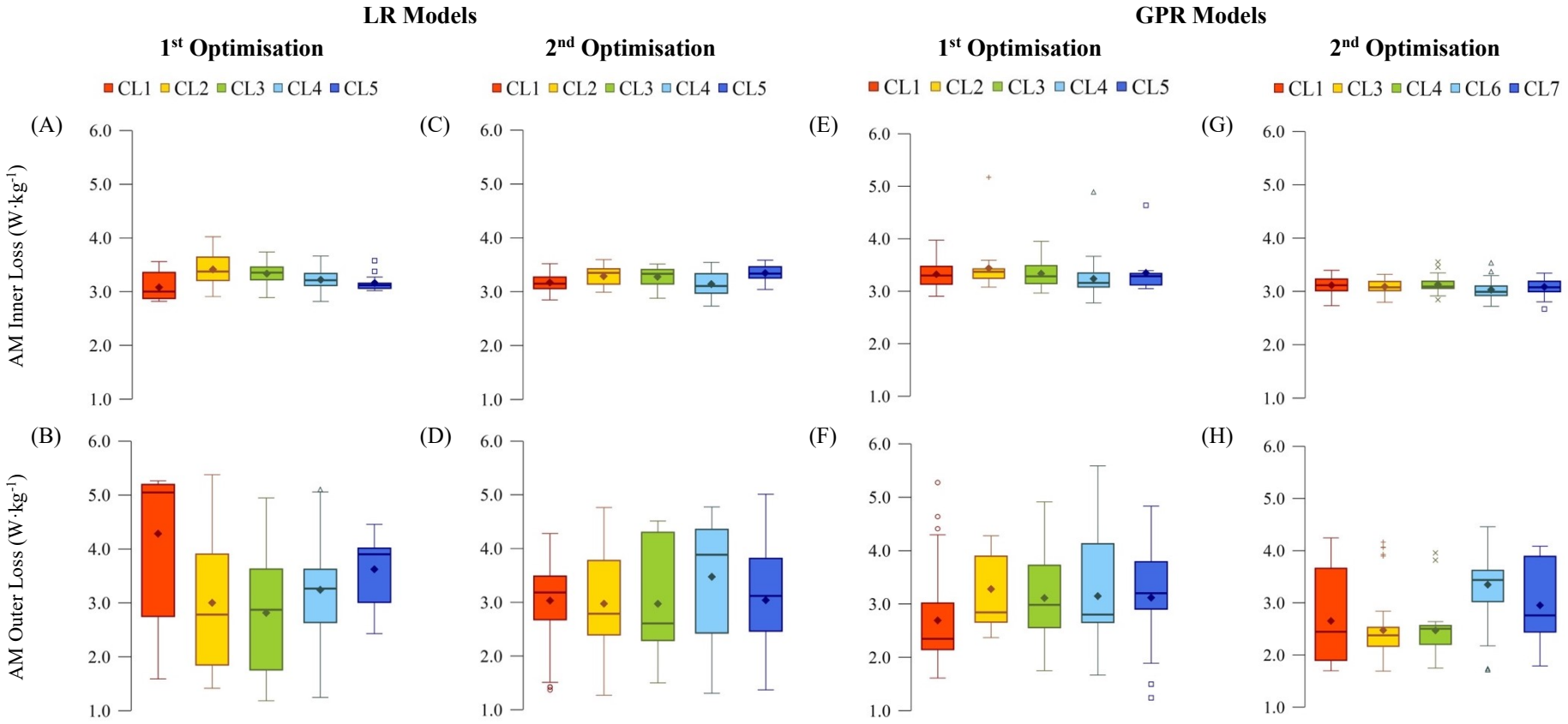


Figure 6-6. AM Loss estimates for the LR and GPR models after the first and second optimisations – five largest clusters. Wide variations in loss for outer validation (bottom row) compared to inner validation (top row). The second optimisation tended to reduce loss variation, and in some cases, the loss estimates were lower, notably for the GPR models. Diamond = Mean; Horizontal Line = Median. Skew indicated by median not being at mid-point.

6.3.4 Optimal models' summary

A breakdown of the optimal LR and GPR models is presented overleaf covering their parameter values and an evaluation of their predictive errors by NCV, MCCV and holdout methods from the second optimisation (Table 6-6 and Table 6-7). Three clustered models are shown in comparison with the non-clustered bagged model averaging the SM series. A new model definition is introduced based on the modal value of the numeric parameters. The modal-defined model outperformed the bagged and clustered models for LR in terms of the MCCV and holdout losses (2.89 $W \cdot kg^{-1}$ and 1.81 $W \cdot kg^{-1}$, respectively). It equalled the best MCCV performance for a GPR cluster model (2.64 $W \cdot kg^{-1}$) but did not match the best GPR holdout performance (2.37 $W \cdot kg^{-1}$). In reviewing the losses, it is clear that in this second optimisation, there were only small differences in performance by the different optimal model definitions used. The greatest disparities can be seen in the holdout losses (1.81–2.07 $W \cdot kg^{-1}$ for LR and 2.13–2.50 $W \cdot kg^{-1}$ for GPR), which is expected because it is a single data set, whereas the other estimates were based on multiple subsamples. In this respect, better holdout performance should be treated with caution.

Table 6-6. LR Models for the second optimisation by grouped observations. The parameters were determined by either the modal peak, averaging (bagging) or clustering (selected clusters shown). Other parameters determined in the first optimisation.

	Modal	Bagged	CL1	CL2	CL3
Group Size	300	300	82	36	23
$\log_{10}\lambda$	-6.48	-7.20	-9.65	-1.44	-10.58
t_{pre} (ms)	1720	1888	2332	1684	1552
t_{post} (ms)	1016	1532	960	1360	1164
$\log_{10}\lambda_{FS}$	7.44	7.36	7.55	6.43	7.48
NCV Loss † ($W \cdot kg^{-1}$)	n/a	3.20	3.03	2.97	2.97
MCCV Loss ‡ ($W \cdot kg^{-1}$)	2.89	2.97	2.96	3.03	3.05
Holdout Loss ($W \cdot kg^{-1}$)	1.81	2.07	1.83	1.85	1.98

† AM Outer Loss. (Mean) ‡ AM Loss. (Aggregate)
NCV = Nested Cross Validation: 10 × 2; MCCV = Monte Carlo Cross Validation: 1000 × 10
Regularisation Method = Ridge; Solver = Least Squares; Alignment = Take-off; No. Basis Functions = 94.

Table 6-7. GPR Models for the second optimisation by grouped observations. The parameters were determined by either the modal peak, averaging (bagging) or clustering (selected clusters shown). Other parameters determined in the first optimisation.

	Modal	Bagged	CL1	CL3	CL4
Group Size	280	280	44	45	26
Basis Function	None	None	None	None	None
$\log_{10}\sigma$	-2.20	-1.93	-1.64	-2.32	-1.65
t_{pre} (ms)	1432	1904	1284	1468	2632
t_{post} (ms)	1208	1732	1212	2440	1336
$\log_{10}\lambda_{FS}$	4.80	2.98	4.57	4.58	4.34
NCV Loss † ($W \cdot kg^{-1}$)	n/a	2.75	2.65	2.47	2.47
MCCV Loss ‡ ($W \cdot kg^{-1}$)	2.64	2.77	2.64	2.66	2.65
Holdout Loss ($W \cdot kg^{-1}$)	2.43	2.13	2.44	2.50	2.37

† AM Outer Loss. ‡ AM Loss.
NCV = Nested Cross Validation: 10 × 2
MCCV = Monte Carlo Cross Validation: 1000 × 10
Kernel Function = Exponential; Standardisation = No; Alignment = Take-off; No. Basis Functions = 102.

6.3.5 Parameter distributions

The numeric parameter distributions show the spread of optimal values across the $\mathrm{SM}_m^{(k)}$ series (Figure 6-7). The clustered models' distributions are also shown within the overall distribution to provide more insight into which aspects of the distribution they represent. The t_{pre} distributions are broad, indicating no strong preference for the time window's start, although for LR, the range is narrower (Figure 6-7A & B). The CL1 model (red), which favours a long time window (2332 ms average, Table 6-6), is the model with the smallest variance in AM outer loss. Of the two clustered models that performed well, CL2 (yellow) is centred on a shorter window (1468 ms) while CL3 (green) covers the longer window (2632 ms, Table 6-7). The distributions for t_{post} indicated a preference for a time window at the narrower end of the range (Figure 6-7C & D) with a close agreement between the two models: 1016 ms for LR (Table 6-6) and 1208 ms for GPR (Table 6-7). In summary, the LR-CL1 model had a time window with a longer period before take-off but a shorter period afterwards. The high-performing GPR models, CL2 and CL3, represented alternate approaches of 'short-long' and 'long-short' time windows, respectively, for the periods before and after take-off.

The roughness penalty, λ_{FS}, for both models had a tall, narrow, bell-shaped distribution indicating that the optimal values were restricted to a short range, centred around $10^{7.44}$ and $10^{4.80}$, respectively, for each model (Figure 6-7E & F). Hence, the LR model appears to benefit from a heavier level of smoothing compared to GPR. Both needed a higher degree of smoothing for optimal performance than GCV had determined (10^2).

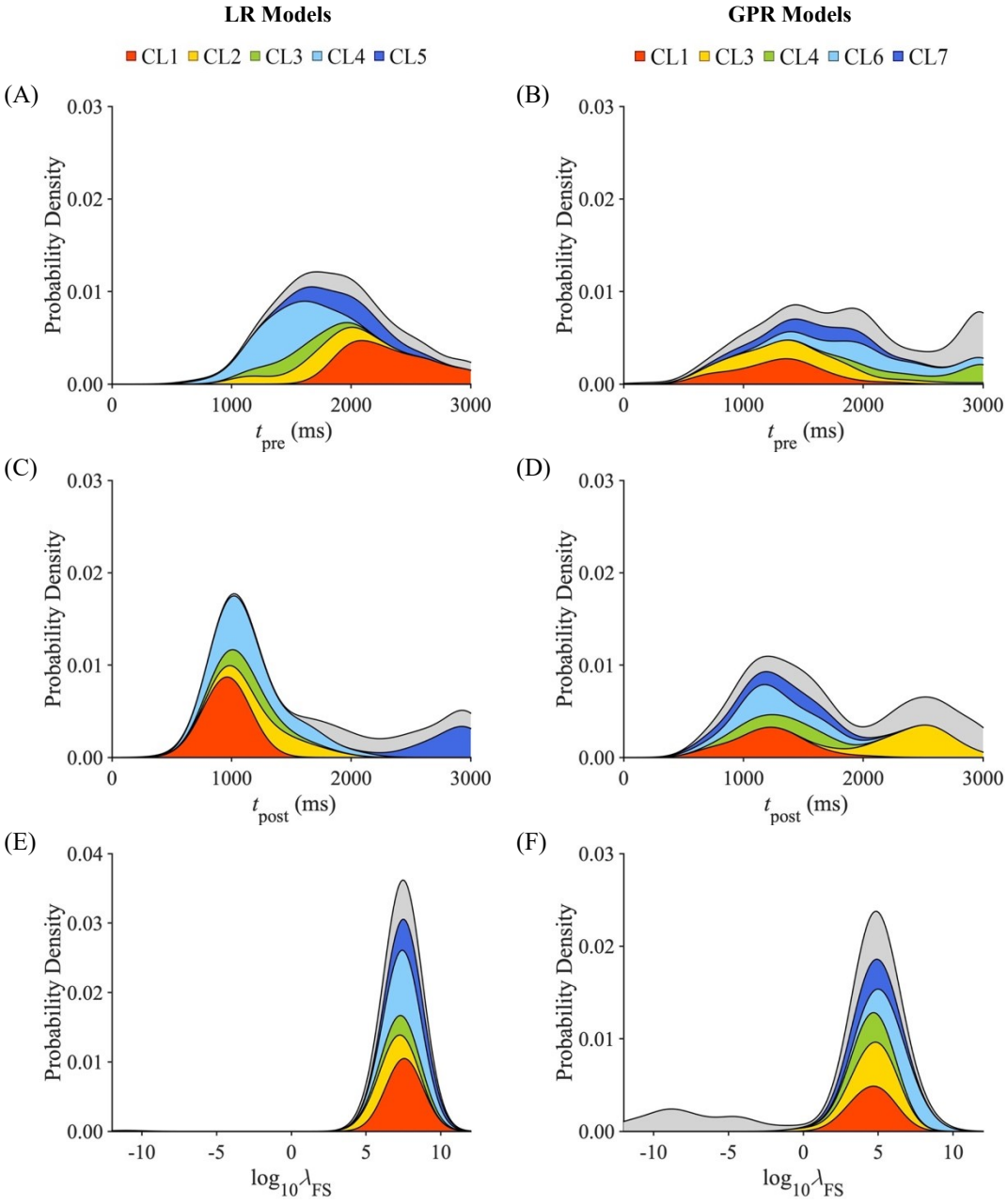


Figure 6-7. Distribution of the data preprocessing parameters for LR and GPR models from the second optimisation: time window, t_{pre} and t_{post}, and roughness penalty, λ_{FS}. The clusters, identified above by colour, can be seen within the overall distributions. The grey areas are the remaining unclustered models.

6.3.6 Bagged Model Plots

Based on the surrogate models bagged across outer folds, the partial plots of the parameter landscape reflect the generalised model's behaviour (Figure 6-8). The plots show a broad range for the time window where the generalised model would be

expected to perform equally well (Figure 6-8A–D). These ranges broadly reflect the distribution ranges of t_{pre} and t_{post}. The roughness penalty plot for LR has a noticeable dip that reflects the bell-shaped distribution above (Figure 6-8E). However, for GPR, the minimum is shallow, well within the level of noise (Figure 6-8F).

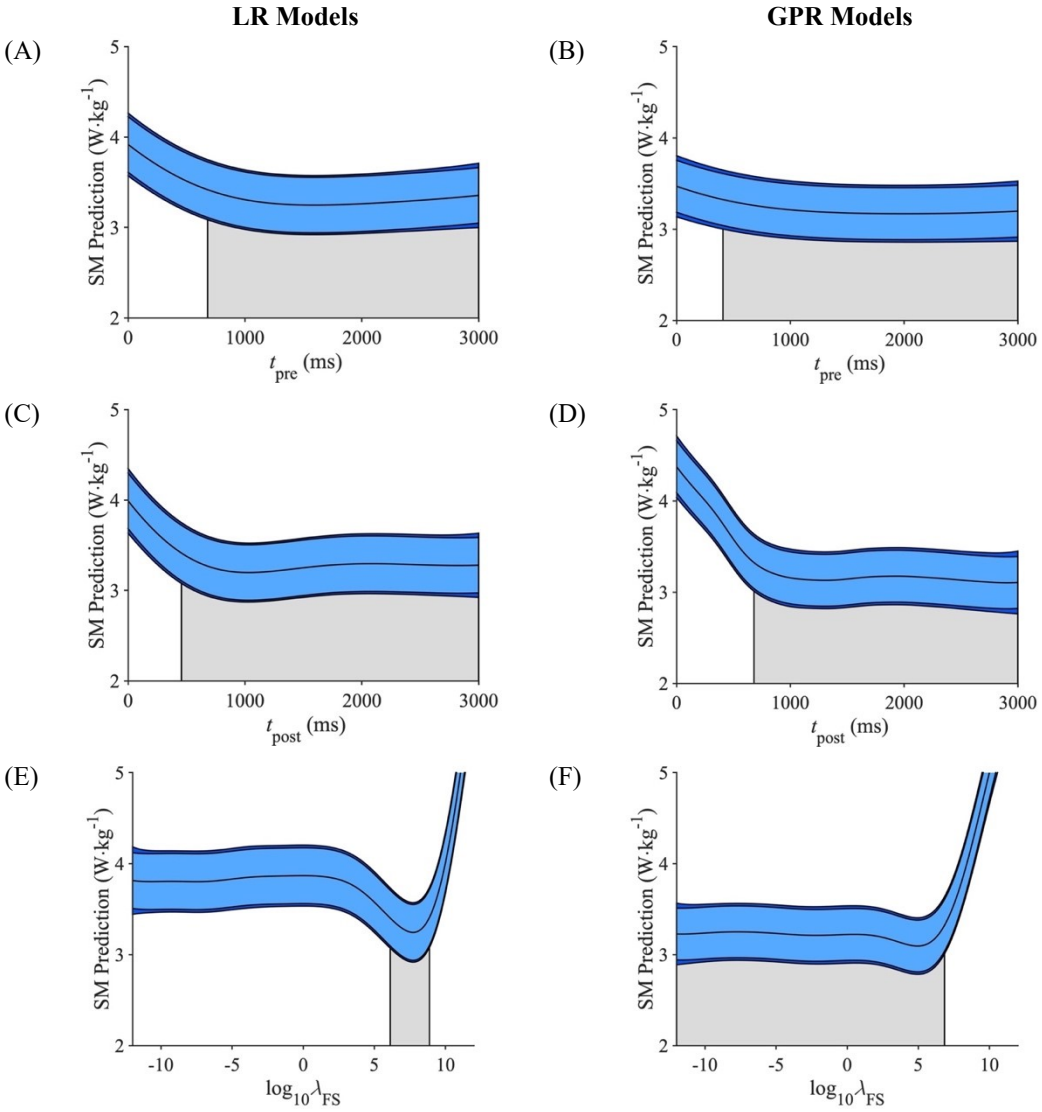


Figure 6-8. Bagged surrogate model plots for LR and GPR models from the second optimisation covering the key data preprocessing parameters defining the time window, t_{pre} and t_{post}, and the roughness penalty, λ_{FS}, determining the level of smoothness. The lightly shaded area underneath the curves indicates the range where equivalent predictive errors are expected due to the inherent level of noise.

6.3.7 Curve registration

Registration was applied to the GPR model alone since it had outperformed LR by a clear margin (2.75 $W \cdot kg^{-1}$ vs. 3.20 $W \cdot kg^{-1}$ for the bagged model; Table 6-7). The optimised models' accuracy based on registered curves was poorer than the non-registered models reported above, achieving at best a predictive error of 2.72 $W \cdot kg^{-1}$ (Table 6-8). There was a strong preference (80%) for using the L_1 landmark on its own (peak power around take-off). There was a change in the preferred basis function to linear and a switch for most models to the Matérn 3/2 kernel. The roughness penalty is also noticeably lower than before with smoothing close to the level needed for a best-fit. The other parameter distributions were otherwise quite similar to those in Figure 6-7. It was not worth pursuing a holdout test because the NCV and MCCV loss estimates fell short of those reported above.

Table 6-8. GPR models optimisation with curve registration during preprocessing that included three different landmark sets and a non-registration set. Only the L_1 landmark was chosen (peak power around take-off).

	Modal	**Bagged**	**CL4**	**CL6**
Group Size	280	280	52	25
Basis Fn	Linear	Linear	Linear	Linear
Kernel Fn	Matérn 3/2	Matérn 3/2	Matérn 3/2	Squared Exp.
$\log_{10}\sigma$	0.29	-1.19	-2.81	-0.27
$\log_{10}\lambda_{FS}$	3.08	2.89	3.04	3.24
Landmark Set	L_1	L_1	L_1	L_1
NCV Loss †	n/a	3.07	2.72	2.80
MCCV Loss ‡	3.00	3.05	3.02	3.07

† AM Outer Loss. ‡ AM Loss.
NCV = Nested Cross Validation: 10 × 2
MCCV = Monte Carlo Cross Validation: 1000 × 10
Standardisation = No; Alignment = Take-off; No. Basis Functions = 100. t_{pre} = 2000 ms; t_{post} = 2000 ms.

6.3.8 Flight times

Based on the modal-defined optimal models (without registration), the validation error dropped sharply for LR and GPR when the time window was extended to include landing as well as take-off (Figure 6-9). The drop was coincident with the actual mean flight time, beginning and ending with the spread of flight times across jumps. The GPR model was more dependent than LR on the acceleration curves, including the landing, as the drop was greater. When the flight time was artificially held constant at 1000 ms, the landing had no discernible effect on the LR model's error, but it increased the error for the GPR model.

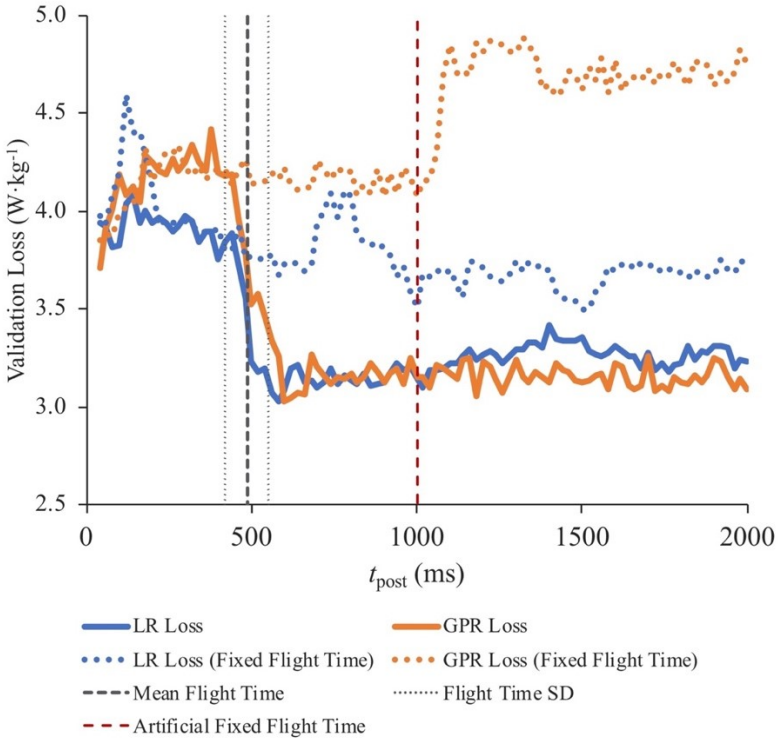


Figure 6-9. Validation loss for increasing t_{post} (t_{pre} = 2000 ms) for LR and GPR models (solid lines) compared to artificial fixed flight time. Model error (solid lines) drops substantially when the time window extends to include the landing for both LR and GPR. With a fixed flight time (dotted lines), LR is unaffected, but GPR error worsens.

6.4 Discussion

The aim of this chapter was to determine the optimal preprocessing parameters for the accelerometer signal. As a result, the GPR models' accuracy improved substantially, with the error falling to the range of 2.5–2.8 $W \cdot kg^{-1}$ from 2.9–3.6 $W \cdot kg^{-1}$ in the previous chapter. In contrast, the LR model benefitted only to a marginal extent, if at all, yielding an error in the range of 2.9–3.2 $W \cdot kg^{-1}$ from 2.8–3.3 $W \cdot kg^{-1}$ in Chapter 5. A full range of comparisons is shown in Table 6-9 based on the different cross validation methods, which reveals large reductions in both models' holdout errors. These improvements were achieved despite switching to using the accelerometer signal to determine take-off, which added a small error of 0.15 $W \cdot kg^{-1}$.

Table 6-9. Comparison of the LR and GPR models between chapters 5 and 6, showing the improvement or otherwise with the inclusion in the optimisation of the data preprocessing parameters.

Model Definition	Model Evaluation	LR Models		GPR Models	
		Ch. 5	Ch. 6	Ch. 5	Ch. 6
Bagging	NCV	3.24	3.20	3.17	2.75
	MCCV	3.25	2.97	3.62	2.77
Best Cluster	NCV	2.84	2.97	2.92	2.47
	MCCV	3.24	2.93	3.24	2.65
Modal	MCCV	n/a	2.89	n/a	2.64
Holdout	Test	3.61	2.07	2.87	2.17

NCV = Nested Cross Validation. MCCV = Monte Carlo Cross Validation. All figures in $W \cdot kg^{-1}$.

This is the first time in this thesis that there has been a clear difference between algorithms following optimisation. Originally, the GPR models were ranked third, behind and LR and SVM, following the initial grid search in Section 5.3.1. However, by extending the optimisation to include the factors governing preprocessing, not just the hyperparameters from the last chapter, the GPR models now stand out from the other algorithms. The success of GPR suggests the Bayesian approach in conjunction with the flexibility of a non-parametric method is advantageous when applied to

FPCA-type models based on acceleration data. GPR is rarely used in the biomechanics literature, if at all, but these results suggest that such models are worthy of further consideration. In light of these results, it is clear that the GPR model alone should be carried forward to the next chapter, along with the data preprocessing method discussed below. Having a single model for further development will help to streamline the investigation further and allow the focus to be fully on the methods in question.

6.4.1 Updated comparison with other studies

The best clustered GPR model (CL4) achieved a validation error of 2.47 $W \cdot kg^{-1}$, according to NCV, and 2.65 $W \cdot kg^{-1}$ for MCCV (Table 6-9). In percentage terms, this is an RMSE of 5.5–5.9%, which is an improvement on the 6.5–7.2% reported in Chapter 5 (Figure 6-10). The GPR model error compares favourably with the benchmark error of 4.6 $W \cdot kg^{-1}$ for the peak power prediction studies (Section 2.3.2). An alternative way to compare model performance is the explained variance, r^2, which is independent of the units involved and does not require domain expertise. The GPR model in this chapter explained 87.4% of the variance in peak power, which is still a long way short of the 98.3% achieved for explaining peak power variance in the VGRF model. However, the refinements introduced in this chapter have brought it closer to the 92.0% obtained for the VGRF jump height model (Table 4-3).

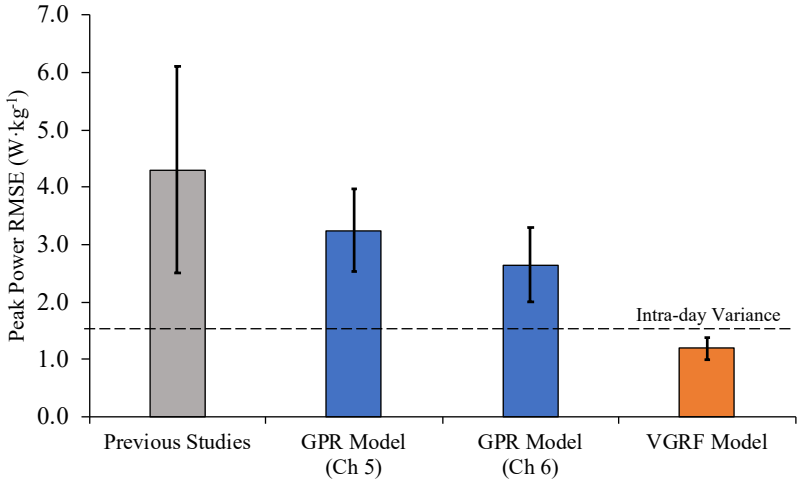


Figure 6-10. GPR models' peak power prediction errors from Chapters 5 and 6 in comparison with previous studies (Ache-Dias et al., 2016; Amonette et al., 2012; Canavan & Vescovi, 2004; Lara et al., 2006; Quagliarella et al., 2010; Tessier et al., 2013). MCCV estimate shown for GPR with error bars based on the SD in validation error between folds. The intra-day variance is averaged from four reliability studies (Cormack, Newton, McGuigan, et al., 2008; Hori et al., 2007; McLellan et al., 2011; Taylor et al., 2012).

6.4.2 Take-off and landing algorithms

The models developed in this chapter depended on a novel algorithm for finding the point of take-off in the accelerometer signal without reference to another data source. The RMSE of 11.9 ms, equivalent to three time intervals, was sufficiently small to limit the rise in the models' predictive error to 0.15 W·kg^{-1}. The landing detection algorithm had a similar error. Combining both algorithms, the RMSE in the flight time was 16.4 ms with respect to the VGRF criterion. This flight time error compares favourably to previous studies using an accelerometer to predict jump height or peak power. Not all such studies report the flight time error directly, but Castagna et al. (2013), Monnet et al. (2014) and Picerno et al. (2011) present errors of 21 ms, 22 ms and 37 ms, respectively. For other studies, the flight time error, Δt_f, can be inferred from the jump height estimates by re-arranging the Newtonian formula for jump height, $h = \frac{1}{8} g t^2$, and taking differences using the quoted random error, Δh, and the mean jump height, $\bar{h}$, to obtain:

$$\Delta t_f = \sqrt{\frac{2}{g}} \left[\sqrt{\bar{h} + \Delta h} - \sqrt{\bar{h} - \Delta h} \right] \tag{6.1}$$

Based on this formula, the inferred flight time errors from other studies were 21 ms (Casartelli et al., 2010), 25 ms (Choukou et al., 2014) and 26 ms (Mauch et al., 2014) and 15 ms (Lesinski et al., 2016). It is clear that the take-off detection algorithm developed in this chapter was more accurate than in previous studies, with one exception, which had a directly comparable error. The two studies with the sensors worn on the lower back had the best and the worst accuracy (Lesinski et al., 2016; Picerno, Camomilla, et al., 2011). All the other studies used the Myotest sensor worn on the hip. For the present study, the algorithm and the additional taping employed to hold the sensor firmly in place may have helped. In practice, this is harder to accomplish with tape for the UB-attached sensor. Undergarments specifically designed to house GPS-tracking sensors may not hold the unit firmly enough for an inertial sensor.

Algorithms have been developed to detect footstrike and push-off in walking and running using wearable sensors, which is similar to jump landing and take-off. Their purpose in most studies was to classify gait phases rather than determine the timing of gait events. In the latter case, the algorithms developed to determine gait-event timing have focussed almost entirely on footstrike, which is akin to jump landing (Maiwald et al., 2015; Mercer et al., 2003; Sabatini et al., 2005; Sinclair et al., 2013). The sensors were attached either on the shank or the heel, placing them close to the impact.

The footstrike-detection algorithms in the literature were more straightforward than those in the present study. The tibial-based algorithms found the minimum or the zero point before the longitudinal acceleration peak (Mercer et al., 2003; Sinclair et al., 2013), not unlike the landing-detection algorithm (Section 6.2.1). The heel-based algorithm used the impact acceleration spike (Maiwald et al., 2015) or the minimum angular acceleration from a gyroscope (Sabatini et al., 2005). No study reporting toe-off timing accuracy could be found. The Myotest algorithm, according to Casartelli et al. (2010), found the time difference between the peaks in its estimate of vertical velocity, which is simpler than the multi-step take-off and landing detections algorithms used in the current study (Section 6.2.1). These new algorithms were refined using the Bayesian optimiser to find the best settings to minimise error, demonstrating how Bayesian optimisation can improve biomechanical algorithms. It

is an expert-led approach, refined with optimisation techniques, that achieved high levels of accuracy compared to equivalent jump-based algorithms.

Alternatively, an ML model might have been able to pinpoint take-off or landing accurately. However, an FPCA-based model would require prior curve alignment, which is the purpose of the algorithm in the first place. This problem may be solved using the impact acceleration spike as the alignment point, but with variations in flight time, the curves associated with take-off would be poorly aligned. Failing that, an LSTM neural network could be trained on the accelerometer signals to predict the chosen alignment point. However, so far as this application is concerned, little more can be gained because the impact on model error was only marginal (Figure 6-4). Errors of 8 ms or less would have no discernible effect on the model. However, for applications where flight time needs to be determined, the take-off and landing detections algorithms may be helpful.

6.4.3 Curve alignments

Aligning the curves at take-off was more conducive to predicting peak power, even though the take-off and landing detection algorithms had similar errors. In general, FPCA-type models are more accurate when the curves are aligned because the FPCs captured the curves' amplitude variances more efficiently with fewer components (Kneip & Ramsay, 2008). Increases of 0.09 $W \cdot kg^{-1}$ and 0.15 $W \cdot kg^{-1}$ in the LR and GPR model's predictive error for a small misalignment error at take-off of 12 ms shows how sensitive an FPCA-based model is to curve alignment. In Chapter 4, the padded VGRF curves were more closely aligned near take-off than the time-normalised ones, which needed registration to improve alignment. In the present investigation, the curves could be aligned at either take-off or landing (landing itself, VGRF > 10 N, or the impact peak). If the curves were aligned at take-off, the FPCs were more attuned to amplitude variance around this point, including when peak power is achieved. At the same time, the FPCs would be less sensitive to amplitude variance around landing as the curves will be phase-shifted due to differences in flight time. Such phase shifts will diminish cross-sectional amplitude variance (Chau et al., 2005). Conversely, when the curves are aligned at landing (landing itself or impact

acceleration peak), the situation will be reversed. Hence, the optimisation revealed a clear preference for aligning the curves at take-off rather than landing (Figure 6-5). The inertial accelerations of the jumping movement were more valuable to the model than the accelerations upon landing. However, according to the GLM analysis, the difference in model error between the two situations was minor compared to other factors (+0.27 $W \cdot kg^{-1}$ for LR in Table 6-3; +0.46 $W \cdot kg^{-1}$ for GPR in Table 6-4). Hence, the inertial accelerations on landing still have relevance to predicting peak power, but not to the same extent as they do around take-off. The relative benefits were weighted more strongly in the GPR model. The GPR model was also less sensitive to large misalignments at take-off, as the Gaussian noise curve revealed (Figure 6-4).

6.4.4 Curve registration

Curve registration was investigated to determine whether it could improve the GPR model's accuracy through better curve alignment. It held out the possibility of warping the time domain to align the curves at both take-off and landing. In theory, time domain warping could nullify the phase-shifting from different flight times. Registration had limited benefits to the VGRF models, correcting for misalignments arising from time-normalisation, more so for jump height than peak power models (Table 4-1). This chapter aimed to determine whether it was an effective technique when applied to accelerometer data.

The results showed that registration did not improve the GPR model's accuracy. The best error was 2.72 $W \cdot kg^{-1}$, which was worse than the 2.47 $W \cdot kg^{-1}$ achieved earlier without registration. In practice, registration was hindered by the considerable variability between the acceleration curves. The only common landmarks that could be identified were based on the pseudo power curves. That was not necessarily an issue because the VGRF models benefited from the zero and peak power landmarks (Table 4-2). In theory, the two peak power landmarks could help to control for variations in flight time as the temporal components described the phase shift. However, the optimisation consistently favoured the first peak power landmark, L_1, alone, rather than combining it with L_2 for reasons discussed below (Section 6.4.5). The

optimisation did not select L_2 on its own because it left curves phase-shifted around take-off, the period which was more relevant to the models.

The problem with the accelerometer data was that L_1 could occur before or after take-off. These large temporal variations effectively prevented take-off and peak power from becoming joint landmarks because landmarks must appear in a consistent order. In comparison, the single peak power landmark in the VGRF curves had combined well with the implicit take-off landmark, ensuring close alignment over a relatively short period when power development reached its peak (Section 4.4.2). The variance in L_1 timing may be attributed to differences in the pseudo power generated during hip and back extension, which may continue after take-off, depending on the hip joints' range of motion and the need to control segmental rotations in flight.

Another factor was the need to apply standardisation to the component scores. The models were much less accurate without it because the temporal FPC scores were up to three orders of magnitude larger than the amplitude FPC scores. However, as seen in the previous chapter, standardisation tends to increase predictive errors (Section 5.4.4). Although registration helped improve accuracy to a limited extent, it could not overcome the handicap of standardisation. Furthermore, registration itself was a heavy computational burden requiring processing in advance of the optimisation over several hours. Consequently, the optimisation had to exclude t_{pre} and t_{post} because the time window had to be fixed to prepare the registered curves. In conclusion, registration was not an effective technique with accelerometer data because the signals were too variable to allow useful landmarks to be identified.

6.4.5 Flight time

The influence flight time had on the models was investigated further by comparing their predictive errors in two conditions: (1) when the flight time was allowed to vary freely (no intervention); and (2) an artificial situation in which the flight time was fixed. The results revealed that the model error only dropped substantially when the flight times were allowed to vary freely (Figure 6-9). These findings demonstrate that the models depend primarily on the flight time rather than the accelerations patterns of landing. It follows that there is no advantage in aligning the curves at both take-off and

landing because the information about the flight time is lost. Registration had attempted to do this when landmarks L_1 and L_2 were jointly specified. It had the advantage of the temporal FPCs retaining information related to the flight time, but the variable positioning of L_1 undermined the flight time correlation. A similar situation would arise with dynamic time warping (Sakoe & Chiba, 1978), which achieves the same outcome as curve registration in a non-functional context and has been used extensively in many applications, including gait analysis (Boulgouris et al., 2004; Helwig et al., 2011; Senin, 2008). It would lead to the same result as it would effectively remove flight time information from the data.

6.4.6 Optimal time window

Determining the optimal time window was one of the aims of this chapter. t_{pre} and t_{post} were measured from the best alignment point, which turned out to be take-off, as discussed above. The optimal values for t_{pre} and t_{post} can be put in context by reference to the resultant acceleration signal, as shown below in Figure 6-11. The acceleration curves have a close correspondence with the more familiar VGRF curve. All the t_{pre} times for GPR (modal, bagged and cluster models, Table 6-7) would be sufficient to encompass the acceleration variance ($t_{pre} \geq 1284$ ms). In fact, the t_{pre} distribution (Figure 6-7B) shows that the time window extended well beyond the point that would appear sufficient to cover the jumping movement. The spread of jump execution times (up to take-off; Figure 6-12) explains why the optimisation resulted in longer time windows (modal time of 1732 ms with a tail extending to 3000 ms). In broad terms, the t_{pre} distribution is shifted to the right of the execution time distribution, so the time window encloses all the inertial accelerations. Therefore, it would be appropriate to set $t_{pre} = 1500$ ms, which is similar to the modal value (1432 ms). It marks the point where the upper tail of distribution substantively comes to an end (Figure 6-12).

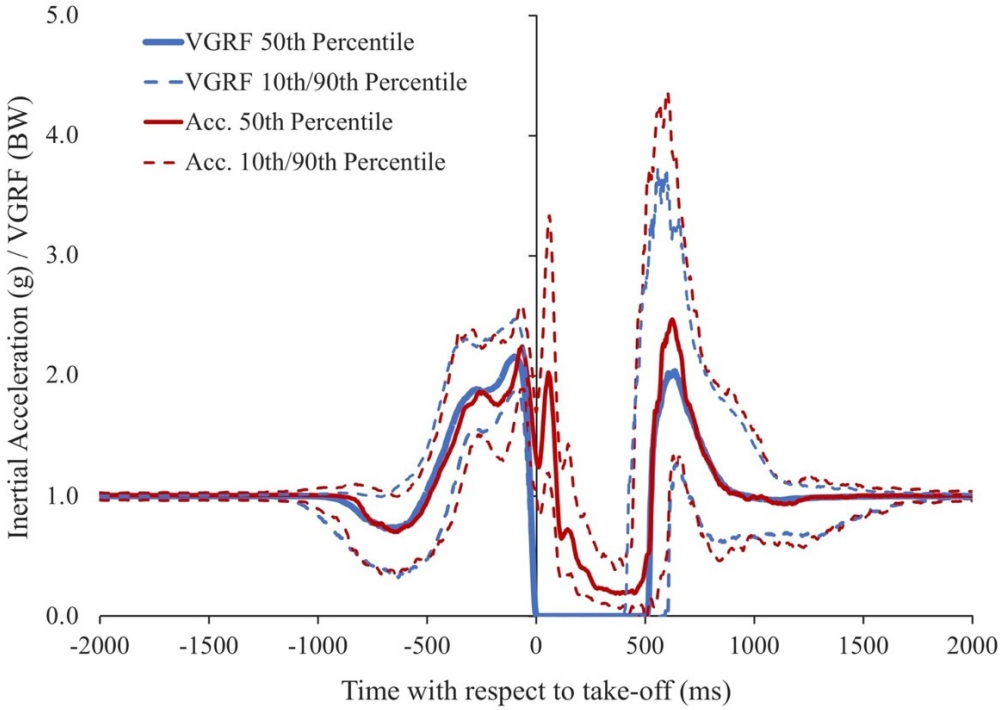


Figure 6-11. Comparison of VGRF and resultant acceleration (LB sensor) for the CMJ_{NA} based on smoothed data. Percentiles are shown to represent spread rather than mean and SD because the distribution at a given time was not normal.

The peaks in the t_{post} distributions (1208 ms for GPR, Table 6-7) correspond to the region where the mean inertial acceleration has begun to stabilise, although some residual variance remains between individual curves (Figure 6-11). Nonetheless, 1208 ms was sufficient for the time window to enclose almost all of the acceleration variance after take-off. This period extends far beyond the landing, with a median of 486 ms and a range extending from 312 ms to 636 ms (Figure 6-13). In conclusion, it would be reasonable for t_{post} = 1200 ms (rounding the modal value), which includes 600–900 ms after the impact of landing when the participant would be regaining their standing posture.

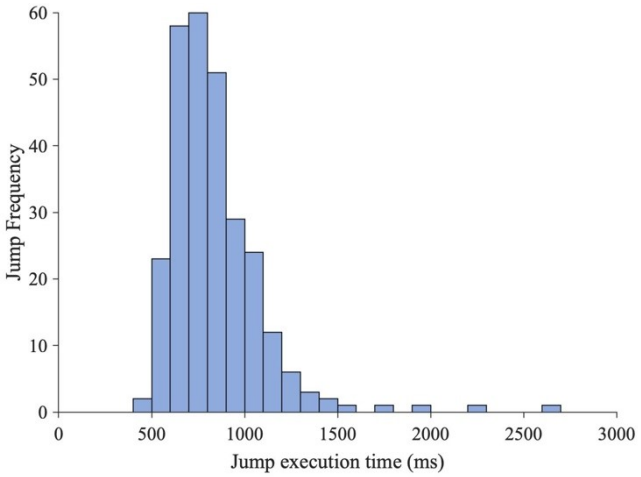


Figure 6-12. Distribution of jump execution times based on VGRF data for the CMJ_{NA} in the training/validation data set.

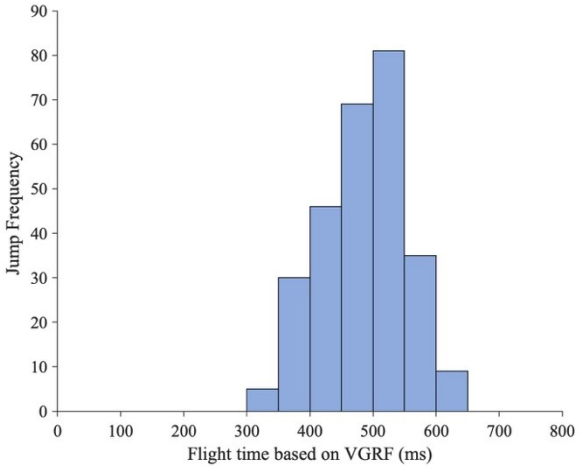


Figure 6-13. Distribution of flight times based on VGRF data for the CMJ_{NA} in the training/validation data set.

6.4.7 Accelerometer signal smoothing

The optimised models preferred a bigger roughness penalty than that determined by GCV so that the curves were smoothed more heavily ($10^{4.8}$ for GPR vs. $10^{2.0}$). Smoothing was more severe for the LR ($10^{7.4}$) despite the linear model seeking to limit complexity itself with regularisation. This preference for heavy smoothing is in line with several activity-recognition studies that have used low-pass filters with low cut-off frequencies, ≤ 3 Hz (Kautz et al., 2017; McGrath et al., 2019; Mlakar & Luštrek, 2017; Zago et al., 2019). Heavier smoothing brought with it a larger fitting error, which

translated into smaller scores for the lower-order FPCs, lessening their relative weight (Figure 6-14A). The scores for the higher-order FPCs, representing minor features, became relatively more important compared to the low-order components representing prominent features (Figure 6-14B). In effect, smoothing re-weights the FPC scores in a similar way to standardisation, which converts the FPCs to Z-scores. However, unlike standardisation, the re-weighting can be regulated by adjusting the roughness penalty accordingly. This re-weighting may more than make up for any additional overfitting from the highest-order FPCs gaining more influence.

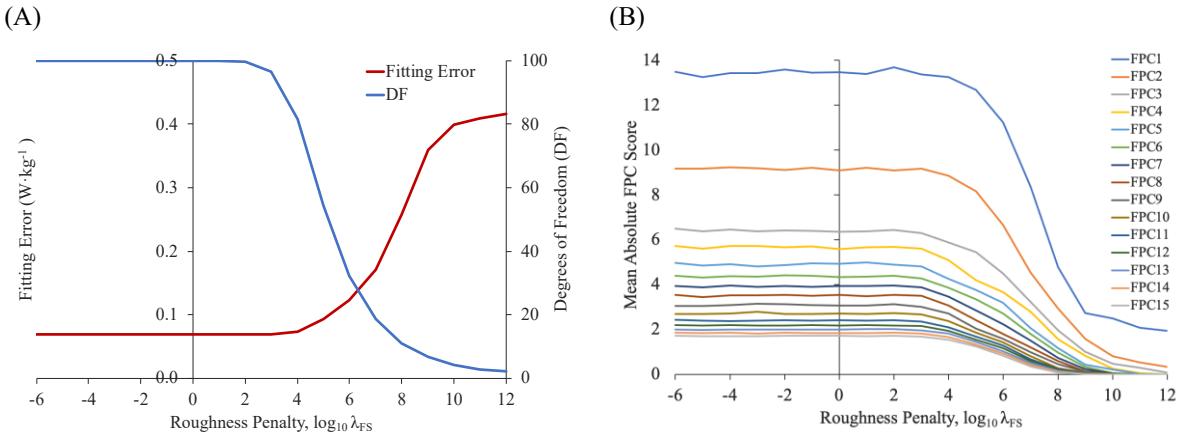


Figure 6-14. Impact of varying the roughness penalty, λ_{FS}, on (A) the fitting error with the raw accelerometer data; the degree of freedom (maximum of 100, the number of basis functions); and (B) the mean absolute FPC scores (absolute magnitude is required because FPC scores are centred around zero, by definition).

6.4.8 Modal and clustered models

The clustered method continued to outperform the bagging approach because it focused on a more tightly defined set of models. The traditional bagging approach took an average of the parameter values across the whole SM series, where outliers can have a disproportionate effect. The mean is favoured in the literature because it tends to provide a more conservative estimate, avoiding overly optimistic errors (Cawley & Talbot, 2010; Hall & Robinson, 2009; Varoquaux et al., 2017). However, the distribution peak can be advantageous when determining the optimal parameter values (Cawley & Talbot, 2010; Krstajic et al., 2014). In those studies, the distributions

formed classical bell-shaped curves because only one parameter was under investigation. However, the current study has used real-world data where multiple factors led to irregular parameter distributions. Clustering was an attempt to compartmentalise the distributions to make interpretation easier. The case of t_{post} illustrates this point where the average is much higher than the modal values (1732 ms vs. 1208 ms in the case of GPR, Table 6-7) due to the rise in probability towards the upper limit (Figure 6-7). Cluster analysis dealt with the situation more effectively by assigning one cluster to 'represent' the high t_{post} models, while the others covered the more common lower range. As the results show, the modal value can be equally effective for GPR and slightly better for LR when considering the MCCV results. The modal value may be considered the most representative as it is drawn from the largest collection of models.

6.4.9 Summary

The models were made self-reliant in this chapter by introducing a take-off detection algorithm, so they no longer relied on VGRF data. This is a vital step in developing a wearable sensor system for predicting peak power in jumping. The accelerometer models were made more accurate by extending the time window on either side of take-off to include landing. Smoothing the data more heavily also reduced predictive errors. The GPR model benefitted much more from these refinements, achieving a predictive error at best of 2.5 $W \cdot kg^{-1}$ (5.5%) compared to only 2.9 $W \cdot kg^{-1}$ for LR. This marked a substantial improvement over the models produced in Chapter 5, where the optimisation focused on the hyperparameters alone. However, the GPR model's accuracy still falls short of the target of 1.75 $W \cdot kg^{-1}$, the intra-day variance.

Further investigations established that the models depended more on the flight time and the inertial accelerations recorded around take-off. Curve registration could not be applied effectively to the accelerometer signals because they were too variable. There may have been misalignments, but no unambiguous landmarks could be found in the acceleration curves. Only landmarks in the pseudo peak power curves could be identified. However, their relative timing with respect to take-off was inconsistent, undermining any benefit that might have arisen from registration.

With a clear differential between the algorithms now, only the GPR model will be carried forward to the next chapter. The optimal settings covering the time window, t_{pre} = 1500 ms and t_{post} = 1200 ms, and the roughness penalty, $\lambda_{\text{FS}} = 10^{4.8}$, will be used in subsequent investigations. The next chapter will examine what additional information can be extracted from the accelerometer signals to reduce the GPR model's error further. This approach will entail generating many more predictors so that feature selection will become a new consideration.

CHAPTER 7. FEATURE SELECTION

7.1 Introduction

The next step in the development of the accelerometer model considers feature selection methods. It answers the fifth research question, *what characteristics of the accelerometer signal are more important to the model in predicting peak power?* In the two previous chapters, the models were based on the resultant accelerometer signal as it provided one time series for each jump. This representation simplified the model, but directional information was lost. Instead, the models could be based on the original triaxial data recording the simultaneous inertial accelerations along the sensor's orthogonal axes. Using such three-dimensional data, the models could distinguish between longitudinal and anteroposterior accelerations in the sensor's local reference frame. Although the body's CM moves primarily in the vertical direction (in the global reference frame), the inertial accelerations recorded by the sensor will shift between the X- and Z-axes in response to the trunk's changing inclination in the sagittal plane. Hence, the orthogonal inertial accelerations will capture more information that may be valuable to the model.

Other potential data representations include curves obtained by differentiating or integrating the acceleration curve. Such curves may yield characteristics related to either the rates of change or other accumulation effects, which are not directly represented by the FPCs of the acceleration curve alone. For instance, the pseudo power curves produced landmarks for the curve registration in the last chapter when other representations produced no unambiguous landmarks. The characteristics of the time derivative curves may have relevance to the peak power model given that the Rate of Force Development (RFD) has been associated with jump height (r = 0.68–0.86) (De Ruiter et al., 2006; Marcora & Miller, 2000; McLellan et al., 2011). Furthermore, training intended to improve peak power also yields increases in RFD (Cormie et al., 2011). This approach has similarities to the convolutions performed by deep neural networks, which reshape and transform the input data into a form more closely related to the outcome variable (Hastie et al., 2009; Ordóñez & Roggen, 2016).

The above representations can generate many more FPCs for the model than in the previous chapter. For instance, running FPCA on three orthogonal signals ('3D signals') would produce three times as many FPC scores. Alternatively, the FPC scores obtained from various derivative and integrated curves could be concatenated to form a long list of predictors for the model. Putting all those components into the model directly without prior selection may lead to overfitting due to redundancy or too much complexity. The most straightforward approach would be to vary the number of retained FPCs. In previous chapters, the number of retained FPCs was held constant while the effects of various model parameters were investigated. Furthermore, the selected components for the VGRF model did not form a contiguous block (Figure 4-4A). Moreover, some FPCs had a strong association with peak power unrelated to the proportion of the curve variance they explained (FPC3: 50.7% vs. 8.6%). Therefore, it would be appropriate to use a selection method in this chapter.

Under the NCV framework, the outer validation loss could guide model selection as the SBC statistic did for the VGRF models (Sections 4.2.5 & 4.2.7). Binary parameters could be defined, specifying predictor selection. However, the multi-dimensional search space would be vast, many times larger than in previous optimisations in this thesis. Furthermore, the last chapter's two-stage optimisation demonstrated that the number of parameters in the optimisation should be minimised. This reflects findings elsewhere that Bayesian methods are not well-suited to large multi-dimensional problems (Eggensperger et al., 2013; Li et al., 2017; Z. Wang et al., 2013). Instead, it would be more efficient to adopt a Monte Carlo approach by generating models with a random selection of predictors (Chen et al., 2018; Draminski et al., 2008). This approach is similar to one taken with the VGRF models, where key components were identified based on their selection frequency (Sections 4.2.7 & 4.3.4). It also reflects the GLM analyses in the previous chapter, which identified key factors influencing the models (Sections 6.2.3 & 6.3.2).

An alternative approach would be to use a thresholding technique, like the one used in Supervised Principal Component Analysis (SPCA) (Bair et al., 2006; Hastie et al., 2009). The SPCA method selects components based on whether their absolute univariate coefficient exceeds a given threshold. A similar threshold-based selection

could be implemented using the optimisation procedure to determine the optimal threshold and other associated parameters.

Therefore, the objectives of this chapter are to:

- Identify the key FPCs responsible for explaining the peak power variance to understand what information is provided to the model;
- Improve the accuracy of the GPR model, where possible, by introducing feature selection for models based on the various data representations above;
- Evaluate the efficacy of the two proposed selection methods based on a comparison of the predictive errors.

7.2 Methods

The feature selection methods implemented were the Monte Carlo and thresholding techniques. The Monte Carlo approach did not involve optimisation but required a statistical model to analyse the model loss. The thresholding technique was subject to optimisation to determine appropriate parameter settings, including the number of FPCs to retain. These selection methods were applied to models based on data in the form of resultant accelerometer signals ('1D models'), the original triaxial accelerations ('3D models') and a set of curves derived from the resultant acceleration based on time derivatives and integration ('multi-1D models'). The modal-defined GPR model from Chapter 6 was used without further optimisation of its parameters.

7.2.1 3D signal preparation

The introduction of a three-dimensional time series did not require any modifications to functional smoothing or FPCA, which processed the data as a vector, $\mathbf{a}_n(t)$. The output from FPCA comprised a set of FPCs for each dimension, tripling the number of predictors available for the model. This trivariate FPCA is similar to bivariate FPCA that has been used elsewhere to investigate rowing technique, among other applications where there is an interaction between variables (Warmenhoven et al., 2017a). However, performing FPCA simultaneously on the three orthogonal functions introduced a correlation between the FPC scores across dimensions. Such correlations

are expected since the orthogonal curves are interrelated: a change in the inertial acceleration vector's direction will bring changes in two or three dimensions.

7.2.2 Generating multi-1D curves

Five new curves were obtained from the resultant acceleration curves ('ACC') using differentiation or integration. These operations were only applied to the resultant curves to limit the number of curves to five rather than 15 in order to keep the data sets manageable. The AM function generated the new curves with each iteration of the random search, incorporating them into the optimisation procedure. The overhead of regenerating the curves was relatively light compared to the computational cost of FPCA and model fitting. The AM function obtained the first- and second-time derivatives ('AD1' and 'AD2', respectively) using functional differential operators (Ramsay & Silverman, 2005). There are no such operators for integration, so a similar procedure was required to the one used to produce the registered power curves (Section 6.2.4).

The AM function converted the acceleration function into a series of points, ten times the number of original basis functions, given that integration amplifies even small errors. It then integrated the time series using the cumulative trapezoidal rule, once to produce the pseudo-velocity curves ('VEL') and a second time to obtain the pseudo-displacement curves ('DIS'). These terms are used advisedly as those curves cannot represent the actual velocity or displacement because the resultant curves ignore the direction of the vectors involved. The pseudo-power time series ('PWR') was then obtained by taking the dot product of acceleration and velocity series. The algorithm then transformed the pseudo velocity, displacement and power time series back into continuous functions so FPCA could process these curves. It then applied FPCA in turn to each of the new curves and then combined the sets of FPCs scores to produce a long predictor vector for each jump.

7.2.3 Monte Carlo selection

The first feature selection method was based on a Monte Carlo approach. A set of Boolean parameters was defined, one for each component, specifying whether to

include the associated FPC in the model. This approach required considerably more parameters than seen previously: for the 1D model, there were 30 parameters (30 FPCs); for the 3D model, 45 parameters (3 × 15 FPCs); and for the multi-1D model, 90 parameters (6 × 15 FPCs). The latter two forms reverted to 15 FPCs per curve to avoid a very large search space of up to 270 parameters. Such a high-dimensional parameter space was too extensive for the optimisation procedure (Section 7.1). Instead, it was more efficient to fit a GLM to the loss using the parameters as categorical predictors. It was a different type of surrogate model from the GP model used previously (Section 5.2.8) and did not require optimisation. The differences in least-squares means could identify the predictors playing a significant role. The model was fitted using the GLMSELECT procedure in SAS with stepwise selection, as in previous chapters (Sections 4.2.5 & 6.2.3). The observations were collected using a random search without convergence in which the FPC selection parameters defined the search space. For the 1D and 3D models, 1000 observations were sufficient (10 × 100), while 2000 were needed for the larger multi-1D model (10 × 200). Once identified, the FPC selection parameters were specified in the Matlab script. The models were then evaluated using the same MCCV design as above and separately for the holdout test. No nested evaluation could be produced for this method.

7.2.4 Correlation threshold selection

For the threshold-based approach to feature selection, the AM function selected FPCs whose absolute pairwise correlation with peak power exceeded a specified threshold. Correlation was preferred as the metric because it would allow direct comparisons with the FPC's explained curve variance. It produced the same FPC rank order as the absolute *t*-statistic employed by the SPCA code library (R. Tibshirani, 2020). It has a high degree of similarity with the standardised univariate coefficient used in SPCA (Figure 7-1), but the SPCA approach requires an additional linear model (Bair et al., 2006).

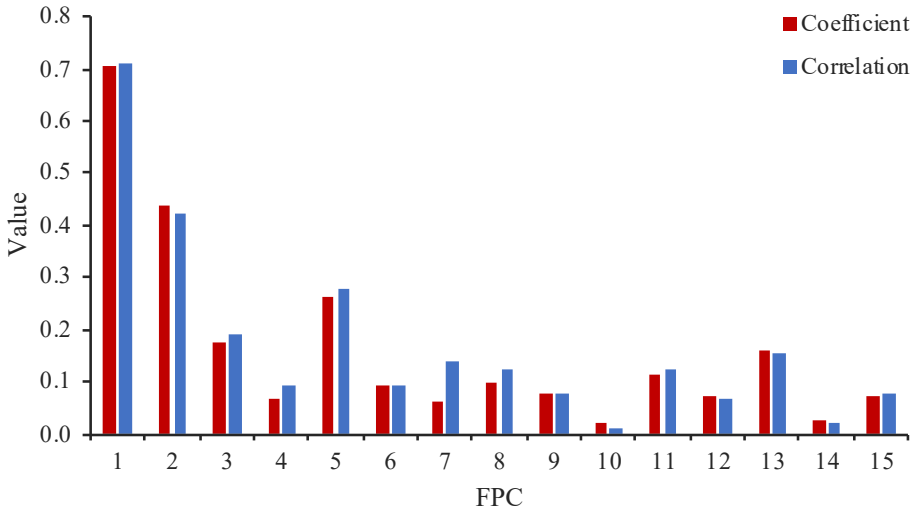


Figure 7-1. FPC score correlation with peak power compared to the standardised univariate coefficient. The absolute value for both statistics is shown. The figures are based on the full training/validation data set.

The correlation threshold was one of four parameters governing feature selection that were subject to optimisation. The other three included the number of retained FPCs from FPCA and two logical conditions determining whether to perform varimax rotation and a signal orientation correction suitable for 3D signals (Table 7-1). The number of retained FPCs was a factor since it defined the pool of components from which to select. The models until this point had used unrotated FPCs, but with the focus on the features themselves, it was appropriate to consider varimax rotation. A logarithmic scale was used for the correlation threshold to allow for more nuanced variations at the lower end of the range where minor differences could have a considerable effect.

Table 7-1. Optimisation parameters for feature selection.

Parameter	Type †	Parameter Values	PSO Bounds ‡
No. retained FPCs	I	{1, …, 30}	[0.50, 30.49]
Varimax	C	{No, Yes}	[0.50, 2.49]
Correlation threshold, r	R	10^{-4} … $10^{-0.5}$	[-4.0, 0.5]
Orientation correction	C	{No, Yes}	[0.50, 2.49]

† Parameter type: C = Categorical (nominal variable); R = Real (continuous), I = Integer (discrete).
‡ Categorical parameters are indexed; real parameters are log-transformed.
Basis Function = None; Kernel Function = Exponential; $\log_{10}\sigma$ = -2.20; Standardisation = No.
Alignment = Take-off; t_{pre} = 2000 ms; t_{post} = 1200 ms; $\log_{10}\lambda_{FS}$ = 4.80; No. Basis Functions = 100.

An orientation correction was introduced to address the differences in the sensor's orientation between individuals due to postural differences (e.g. lumbar curvature), soft tissue differences (e.g. the hollow around the spine in the lumbar region), or minor placement errors. Variations in the sensor's orientation could be detected and measured during the brief period of quiet standing before the jump when gravity was the only measured inertial acceleration. The acceleration vector was averaged over the first ten samples (40 ms) when the participant stood still, before they heard the instruction to jump. The recording was started before the command was given (Section 3.2.2). The whole time series was then rotated using standard matrix methods to align the initial acceleration vector with the vertical (-1, 0, 0). This direction was chosen to minimise the rotational angle as the sensor was attached to the body with the X-axis pointing upward. The Y-axis represented the mediolateral, and the Z-axis pointed in roughly the anterior direction. The mean rotation angle was 16.4° (6.4°–31.4°, 90% CI).

The optimisation used the same nested 10 × 2 design, as in previous chapters, with each observation evaluated using 20 × 2 MCCV. There were fewer observations and models than before (I = 200, M = 10), as there were only four parameters, two of which were binary. The optimised models were then evaluated with 1000 × 10 MCCV and then with the holdout test set.

7.3 Results

7.3.1 Monte Carlo selection

The GPR model based on the resultant (1D) acceleration curves achieved the lowest prediction errors compared to the other larger models (MCCV and holdout loss; Table 7-2). The Monte Carlo method selected a model with 12 out of the 30 components available. According to the corresponding GLM, selecting these components could account for 94% of the variance in the GPR model's predictive error. This proportion was much higher than the 3D and Multi-1D models (68% and 65%, respectively), but GLM's prediction of the GPR model's error was comparable (0.24–0.28 $W \cdot kg^{-1}$).

Table 7-2. GPR models' predictive error based on Monte Carlo feature selection.

	1D	3D *	Multi-1D
GLM Total Partial ω^2	94%	68%	65%
GLM RMSE ($W \cdot kg^{-1}$)	0.28	0.24	0.28
No. Selected FPCs §	12 / 30	17 / 45	17 / 90
MCCV Loss ‡ ($W \cdot kg^{-1}$)	2.73	2.98	2.93
Holdout Loss ($W \cdot kg^{-1}$)	2.58	2.36	2.58

† AM Outer Loss. ‡ AM Loss. § FPC selected if its correlation with peak power exceeds this threshold.
MCCV = Monte Carlo Cross Validation: 1000 × 10. No NCV estimate can be made using this approach.
1D = Resultant Acceleration Signal; 3D = Orthogonal Acceleration Signal;
Multi-1D = Combined set of 6 Resultant Curves: ACC, AD1, AD2, VEL, DIS, PWR.
Basis Function = None; Kernel Function = Exponential; $\log_{10}\sigma$ = -2.20; Standardisation = No.
Alignment = Take-off; t_{pre} = 2000 ms; t_{post} = 1200 ms; $\log_{10}\lambda_{FS}$ = 4.80; No. Basis Functions = 100.
Varimax = No. * Orientation Correction = Yes.

The FPCs selected for the resultant signal included ACC1–ACC11 and ACC13. The first two components explained the overwhelming proportion of the variance in peak power. ACC1 accounted for almost two-thirds (66.1%), ACC2 contributed 19.4%, while the remaining components explained $\leq 3.4\%$ each. With each successive component, the model explained variance monotonically decreased along with the curve explained variance. No FPC stood apart from that trend. It was evident that much of the curve variance for the higher-order FPCs was of little or no value to the peak power model.

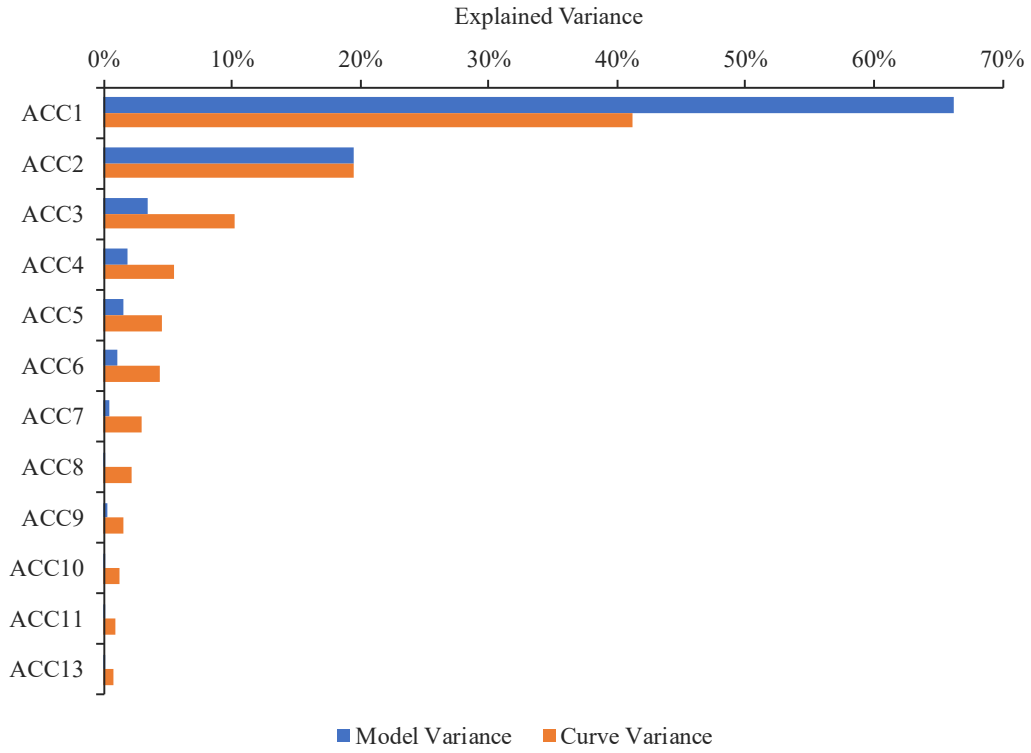


Figure 7-2. 1D components explaining the variance in the GPR model's prediction of peak power (Model Variance) compared to the variance explained in the acceleration (ACC) curve (Curve Variance). The variances relate to the unrotated FPCs.

For the 3D signal, all the X-axis components were included (ACC1X–ACC15X) and the first two FPCs for the Z-axis (ACC1Z and ACC2X). The X-axis pointed in approximately the vertical direction, while the Z-axis was directed in generally the anteroposterior direction. No FPCs were chosen for the Y-axis, which lay on roughly the mediolateral axis. The first components in the X and Z axes, ACC1X and ACC1Z, made the two largest contributions (44.8% and 5.1%, respectively). The other components' contributions were relatively small: 15 out of the 17 FPCs explained $\leq$ 2.2% of the loss variance. In contrast to the 1D model, the explained variance of the X-axis components did not decline so quickly, showing that the contributions from 3D FPCs were not so narrowly spread.

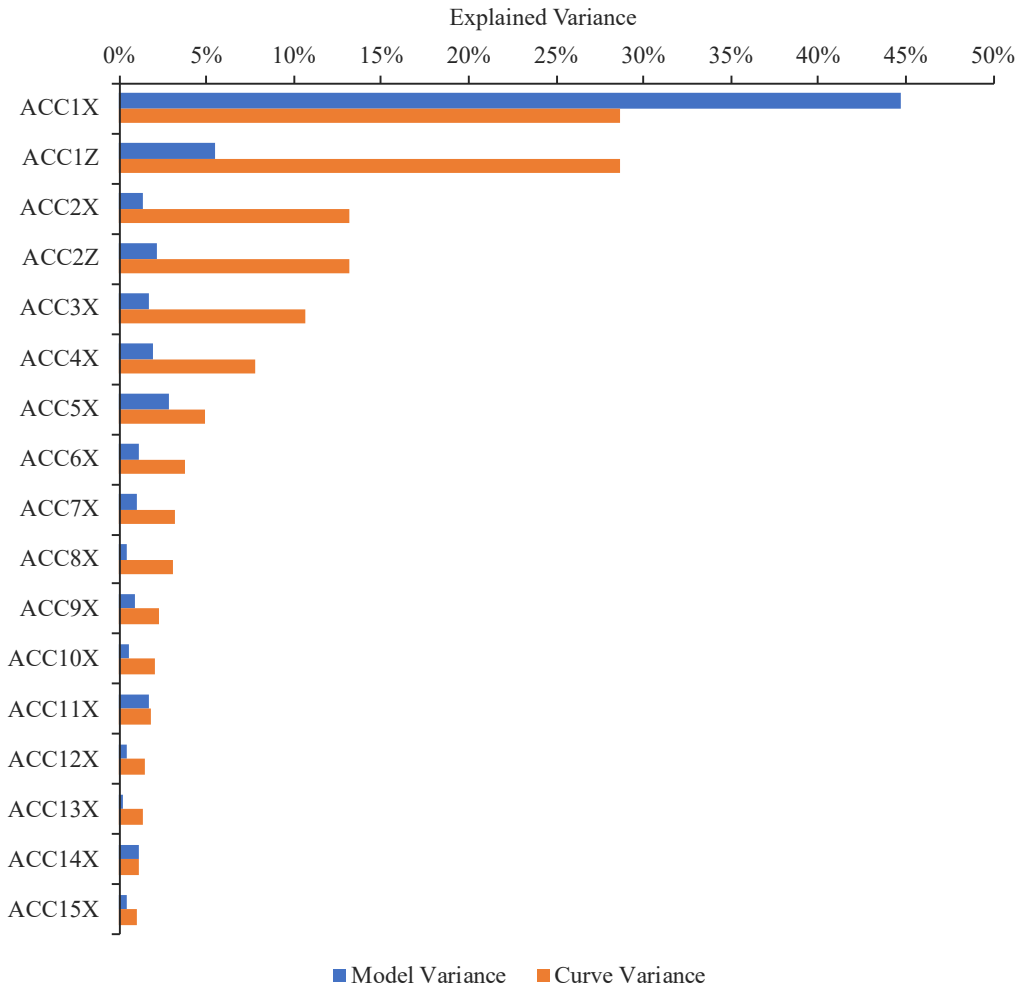


Figure 7-3. 3D components explaining the variance in the GPR model's prediction of peak power (Model Variance) compared to the variance explained in the acceleration (ACC) curve (Curve Variance). All X components were selected, but only two Z components. Curve explained variance is the same across dimensions. The variances relate to the unrotated FPCs.

The features selected for the multi-1D model included only those from the acceleration, pseudo-velocity and pseudo-power curves (Figure 7-4). ACC1, ACC2 and PWR1 explained 21.6%, 25.3% and 5.1% of the AM loss variance, respectively, accounting for 51.9% altogether. The pseudo-curve components, VEL2 and PWR2 were also prominent, explaining 2.7% and 2.1%, respectively. ACC3 split them in the ranking order, contributing 2.2%. The other components accounted for the remaining 6.0%, where each contributed $\leq 1.6\%$. Notably, ACC1 and ACC2 accounted for a smaller proportion of the variance than the 1D model with only the acceleration curves.

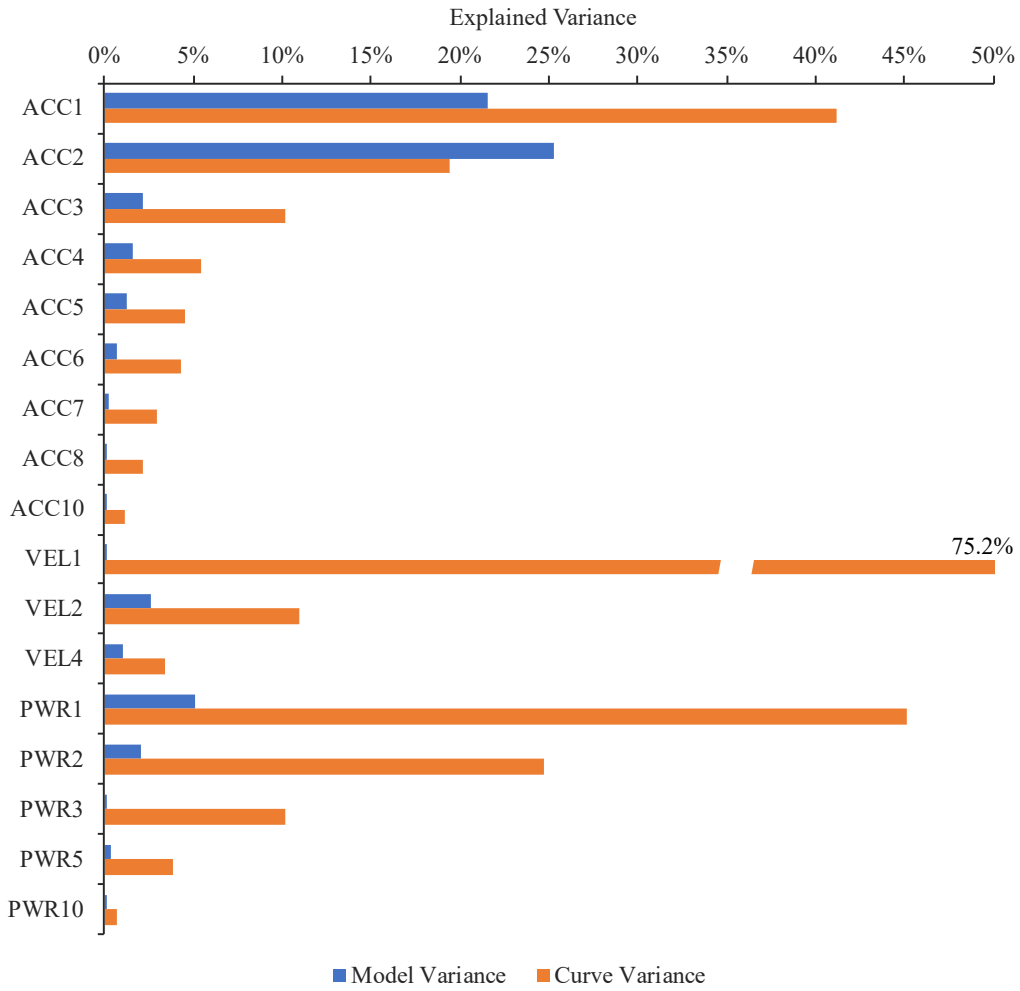


Figure 7-4. Multi-1D components explaining the variance in the GPR model's prediction of peak power (Model Variance) compared to the variance explained for the acceleration curve (ACC), pseudo-velocity curve (VEL) and the pseudo-power curve (PWR) – (Curve Variance). The explained variance relates to the curve in question. The variances relate to the unrotated FPCs.

7.3.2 Functional Principal Components

The first four FPCs for the resultant accelerometer signal are shown in Figure 7-5. The first component, ACC1, represents the temporal variance in the curve predominantly, describing both the variation in the braking and propulsion phases and the variation in flight times (Figure 7-5A). ACC2 primarily represents the variance in the landing impact peak (Figure 7-5B). ACC3 focuses on the variation during the braking phase when the body slows its downward countermovement (Figure 7-5C). ACC4 mainly describes variation in a characteristic peak after take-off, which is more prominent in jumps producing more peak power (Figure 7-5D).

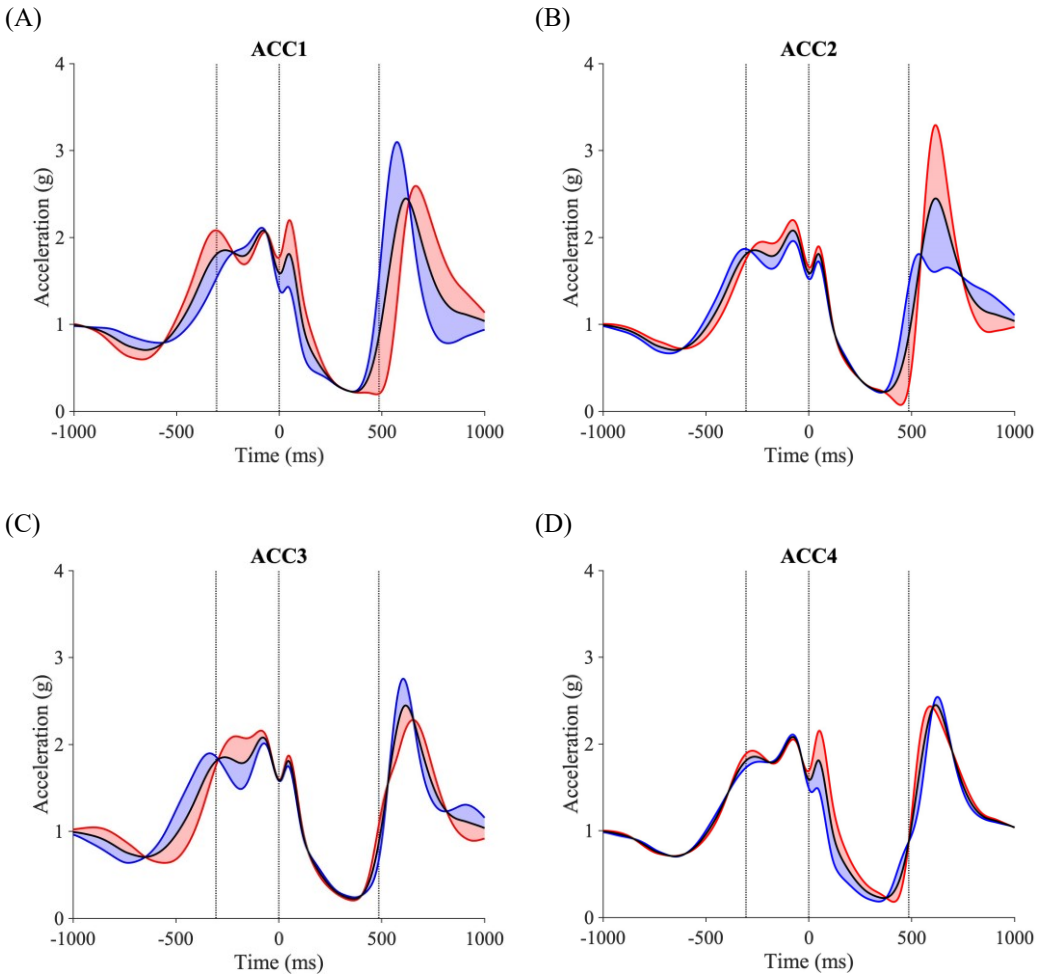


Figure 7-5. Prominent resultant acceleration FPCs (unrotated) correlated with peak power: cross-sectional mean curve (black line); variation correlated with higher peak power (red); variation correlated with lower peak power (blue); vertical dotted lines indicate the mean start of the propulsion phase (-305 ms), take-off (0 ms) where the signals are aligned, and mean landing time (+484 ms). The variances represent one standard deviation. (A) ACC1 indicates temporal variance – jump execution time and flight time; (B) ACC2 indicates variance in impact acceleration; (C) ACC3 indicates variance in the braking and propulsion phases; (D) ACC4 indicates variance in the second spike after take-off.

Selected FPCs from the Multi-1D curves are shown in Figure 7-6. The pseudo-velocity plots show higher velocities for jumps producing greater peak power (Figure 7-6A) and a delayed but faster rise in velocity in the first half of the propulsion phase (Figure 7-6B). PWR1 describes the variation in prominent peak immediately following take-off (Figure 7-6C). The curve's characteristic shape and modality can change depending on the FPC score.

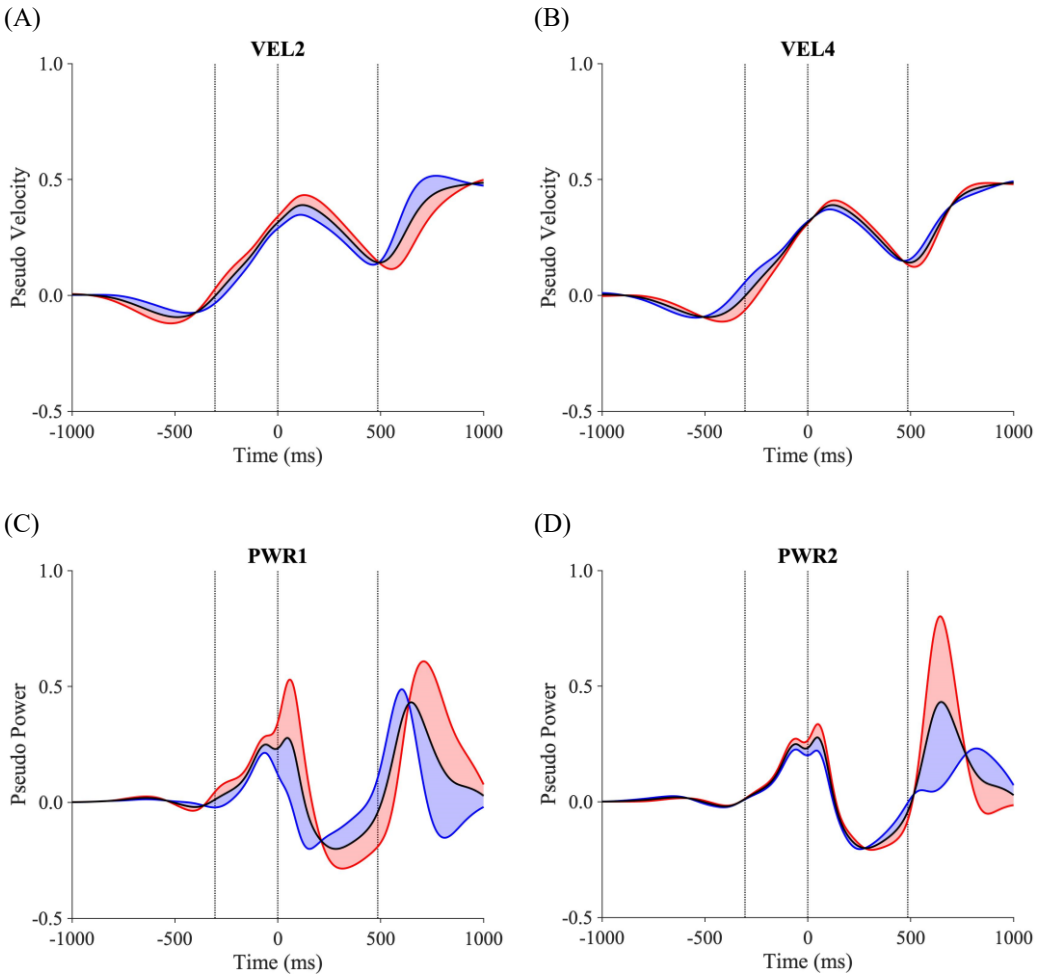


Figure 7-6. Resultant FPCs (unrotated) for pseudo velocity and power that are strongly correlated with peak power: mean curve (black line); variation correlated with higher peak power (red); variation associated with lower peak power (blue); vertical dotted lines indicate the mean start of the propulsion phase, take-off and landing. The variances represent one standard deviation. Vertical axes have been re-scaled for clarity – units are arbitrary. (A) VEL2 indicates variance in velocity throughout; (B) VEL4 indicates amplitude variance (lowest to highest); (C) PWR1 indicates variance peak pseudo power immediate after take-off and on landing; (D) PWR2 indicates variance in landing.

7.3.3 Correlation threshold selection

The GPR model based on the resultant (1D) representation also achieved the lowest predictive errors using the threshold-based selection, using either modal or bagged definitions (Table 7-3). The modal-defined selection included many more components than the bagged selection, achieving lower MCCV and holdout losses (2.62 $W \cdot kg^{-1}$ and 2.14 $W \cdot kg^{-1}$, respectively). These figures were slightly lower than the

corresponding errors using the Monte Carlo selection. The clustered models had predictive errors that were generally slightly higher than those reported here.

Table 7-3. Comparison of GPR model loss (Chapter 6 modal-defined model) for three representations of the accelerometer signal using optimised, correlation-based FPC selection.

	Modal			Bagged		
	1D	**3D**	**Multi-1D**	**1D**	**3D**	**Multi-1D**
Retained FPCs [a]	29/30	13/15	14/15	13/30	12/15	13/15
Varimax	Y	N	Y	Y	N	Y
$\log_{10} r_T$ §	-3.98	-4.00	-3.98	-2.77	-3.83	-3.28
Re-Orientation *	n/a	Y	n/a	n/a	Y	n/a
NCV Loss † **(W·kg^{-1})**	n/a	n/a	n/a	2.62	3.05	2.81
MCCV Loss ‡ **(W·kg^{-1})**	2.62	3.16	2.79	2.69	3.12	2.82
Holdout Loss (W·kg^{-1})	2.14	2.72	2.37	2.55	2.71	2.40

† AM Outer Loss. ‡ AM Loss. § FPC selected if its correlation with peak power exceeds this threshold.
* Re-orientation correction to vertical for 3D signal. [a] Retained FPCs shown out of total available.
NCV = Nested Cross Validation: 10 × 2. MCCV = Monte Carlo Cross Validation: 1000 × 10.
1D = Resultant Acceleration Signal (up to 30 FPCs); 3D = Orthogonal Acceleration Signal (up to 15 FPCs);
Multi-1D = Combined set of 6 Resultant Curves (each up to 15 FPCs): ACC, AD1, AD2, VEL, DIS, PWR.
Basis Function = None; Kernel Function = Exponential; $\log_{10}\sigma$ = -2.20; Standardisation = No.
Alignment = Take-off; t_{pre} = 2000 ms; t_{post} = 1200 ms; $\log_{10}\lambda_{FS}$ = 4.80; No. Basis Functions = 100.

There was an advantage in performing varimax rotation of the FPCs for the 1D models, but not for the 3D model. The correction for sensor orientation for the 3D model was preferred. The optimisation produced models with close to the maximum number of FPCs permitted (29/30 FPCs for 1D; 13/15 for 3D; and 14/15 for multi-1D under the modal definition). Very few FPCs were excluded based on the correlation threshold, whose optimal level was set to close to the lower search bound of 10^{-4}. Across all models, between 65% and 85% of FPCs were chosen every time in 1000 iterations of MCCV. The lowest probability of an FPC being selected in any one model ranged from 0.987 to 0.997. Hence, the number of retained FPCs dictated the selected features, since almost all components from that list were chosen.

For the resultant acceleration model, the loss drops initially as the number of retained FPCs increases, and then crucially, it flattens out with no subsequent rise (Figure 7-7A). There was a range of values (≥ 11 FPCs) where the loss was essentially the same, given the noise level. A slight depression leads to the strongest preference for 29 FPCs, but there are other peaks at 11 and 22 retained FPCs, reflecting slight variations between individual SM plots that are averaged out in the bagged SM plots (Figure 7-7C). There was a marginal advantage in using varimax, but the average gain (-0.09 $W \cdot kg^{-1}$) was less than the noise level (Figure 7-7B). The aggregate correlation threshold plot is almost flat, suggesting no particular advantage in how selective the model should be based on the FPC score's correlation with peak power (Figure 7-7C). However, the threshold parameter distribution reveals peaks at low or high thresholds (Figure 7-7F).

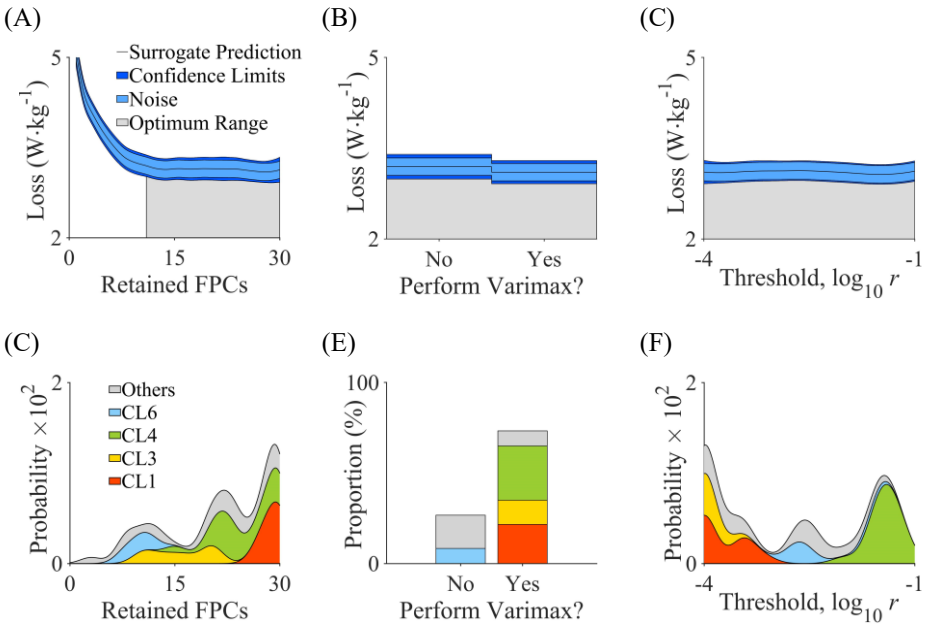


Figure 7-7. Feature selection parameters for a resultant acceleration signal based on a correlation threshold with peak power. SM plots (top row) and parameter distributions (bottom row) with the cluster distributions identified by colour. CL*n* identifies the cluster.

The corresponding plots for the 3D models have similarities with the resultant model for the number of retained FPCs (Figure 7-8A & E) and the correlation threshold (Figure 7-8C). The preference was for a low threshold, meaning that all FPCs were included in the model (Figure 7-8D). There was a strong preference for no varimax rotation, which yielded a gain of 0.19 $W \cdot kg^{-1}$ (Figure 7-8B). The choice for the signal orientation correction was unanimous, which offered an average reduction in error of 0.13 $W \cdot kg^{-1}$ (Figure 7-8D & H).

The equivalent plots for the multi-1D models show that the loss for increasing the number of retained FPCs does not flatten out in the same way but continues to drop, culminating in a strong preference in the parameter distribution for 13 FPCs (Figure 7-9A & D). Varimax made virtually no difference overall to the loss such that the proportional split is close to even (Figure 7-9B & E).

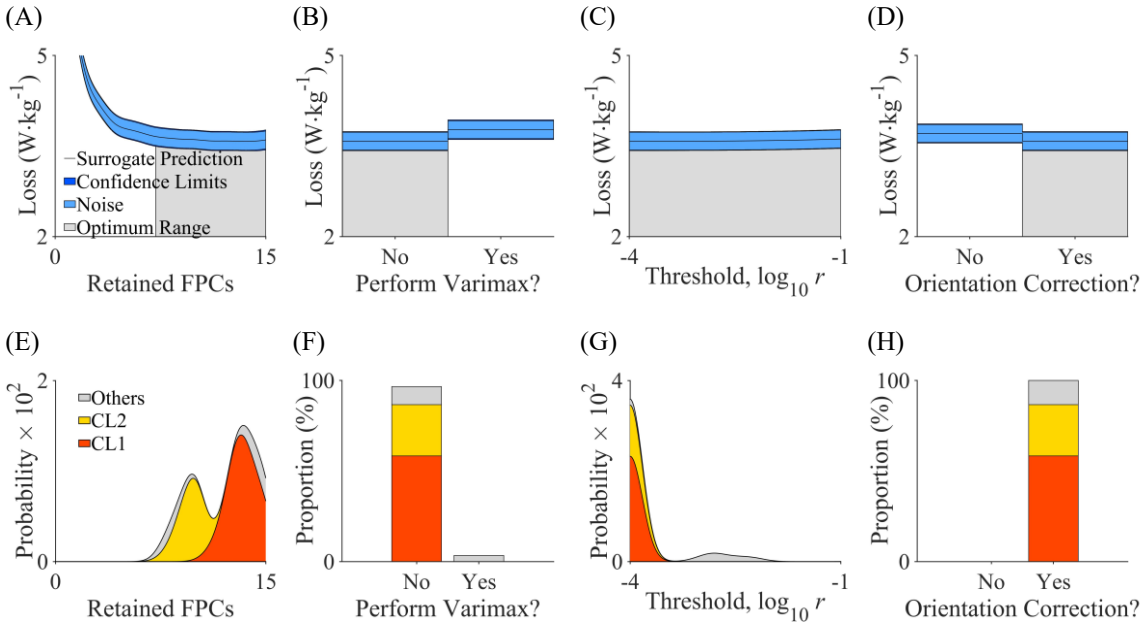


Figure 7-8. Feature selection parameters for orthogonal acceleration signals based on a correlation threshold with peak power. SM plots (top row) and parameter distributions (bottom row) with the cluster distributions identified by colour.

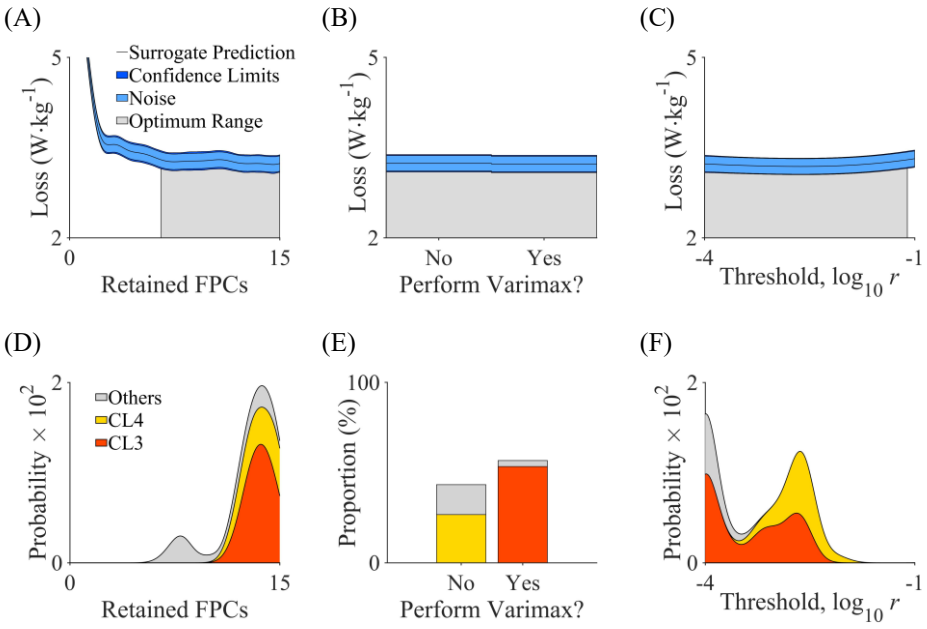


Figure 7-9. Feature selection parameters for a set of derived resultant curves based on a correlation threshold with peak power. SM plots (top row) and parameter distributions (bottom row) with the cluster distributions identified by colour.

7.4 Discussion

This chapter aimed to investigate whether further improvements in the GPR model's accuracy could be achieved by selecting FPCs from a broader list of characteristic features. Feature selection achieved only a fractional reduction in the predictive errors reported in the previous chapter and only when using the resultant acceleration (Table 7-4). Using FPCs obtained from the derived curves also yielded slightly less accurate models. Developing models on the three triaxial accelerations produced the largest errors of all the feature representations considered in this chapter. The reduction in the validation RMSE of 0.02–0.13 $W \cdot kg^{-1}$ for the 1D signal representation (depending on the estimator used) was much less than the noise level (0.38 $W \cdot kg^{-1}$). Hence, there was no meaningful improvement in the GPR model. Nevertheless, the investigation revealed the features in the accelerometer signal that were important to the model.

Table 7-4. Comparison of the best GPR models from chapters 5, 6 and 7.

Model Definition	Model Evaluation	GPR Models		
		Ch. 5	Ch. 6	Ch.7 †
Bagging	NCV	3.17	2.75	2.62
	MCCV	3.62	2.77	2.69
Modal	MCCV	n/a	2.64	2.62
Holdout	Test	2.87	2.17	2.14

NCV = Nested Cross Validation. MCCV = Monte Carlo Cross Validation. All figures in $W \cdot kg^{-1}$.
† Resultant Acceleration signal representation using the correlation threshold method (Table 7-3).

The selection procedure based on the correlation threshold was more effective than the Monte Carlo method, but its optimal correlation threshold was so low ($10^{-4.00}$–$10^{-2.77}$) that almost all components were selected with each subsample (probability 0.987–0.997). The Monte Carlo approach was more parsimonious, and perhaps for that reason, the models were slightly less accurate. It appeared that the GPR model could accommodate large numbers of predictors even though many explained only a small fraction of the peak power variance. Adding more FPCs did not see a rise in the predictive error. Hence, feature selection with FPCA was only a minor consideration in this study.

7.4.1 Key features of the acceleration curves

The resultant accelerometer signal, chosen initially to simplify the analysis (Chapters 5 & 6), turned out to be the best representation of the data. None of the other representations, either in three dimensions or as transformed curves, could provide more useful information to the model. The resultant-based models had only a marginal advantage, but they were simpler and therefore more valuable. The plots of the resultant ACC components (Figure 7-5) also offered the best insight into what patterns of inertial acceleration explained variance in peak power. In this respect, the more discriminating Monte Carlo selection method was helpful.

The ACC1 component represented the temporal variance, including the variation in flight time, the importance of which was first identified in the previous chapter. It

accounted for two thirds of the peak power variance, making it the dominant factor. Similarly, flight time was the metric that underpinned the simpler peak power predictive equations (Section 2.3.2). What the FPCA-based approach was able to do, however, was to add other characteristics that helped to refine the model's predictions. As a result, predictions based on the FPCA-models were more accurate.

The second component, ACC2, represented variations before take-off. This region of the acceleration curve resembled the VGRF curve closely, indicating that the broadly speaking, the sensor was following a similar trajectory to the body's CM. The acceleration curve had a peak before take-off, which appears to reflect the VGRF's FPC3 that was crucial in peak power production (Figure 7-5B, c.f. Figures 4-3C and 4-4A). The acceleration curve also had a characteristic third peak after take-off, which ACC4 mainly describes (Figure 7-5D). This post take-off peak tended to be more prominent in jumps producing more peak power. It was mainly due to a sharp rise in the Z-axis acceleration (approximately the posterior direction), reflecting hip and or lower back extension that appears to have continued after take-off.

ACC2 also showed that more powerful jumpers experience a higher inertial acceleration on landing, a situation that resembles the absorption phase of a drop jump. Bobbert et al. (1987a) showed that the peak negative power absorbing the impact rose proportionately with drop height. The authors reported that the absorption phase duration did not change in response to higher impact velocities. Hence, the kinetic energy was absorbed at a faster rate, requiring more negative power (Bobbert et al., 1987b). In the current study, the impact deceleration was not necessarily related to the flight time because ACC2 was independent of ACC1. This observation suggests that the landing can be 'soft' (a slower deceleration) or 'hard' irrespective of the impact velocity.

The ACC plots revealed other similarities with the VGRF curve, demonstrating the value of using the LB attachment site. Although the sensor may not have moved in train with the body's CM, its movement served as a useful proxy. The plots also highlighted valuable insights into the biomechanics of the CMJ_{NA}, demonstrating the value of using FPCA as the feature extraction method. FPCA captured crucial aspects of the movement that were relevant to the performance outcome, as it has done in many

other applications (Section 2.5.2). However, the information present in the accelerometer signal may not be sufficient to predict peak power with sufficient accuracy (Section 9.4.3). More generally, however, this investigation demonstrates that machine learning models can identify patterns in the data that a researcher may not at first consider. ACC1 and ACC2, the two most important features, both related to events occurring after take-off, a finding that was not anticipated at the outset of this research.

7.4.2 Differentiation and integration curves

Obtaining curves derived from the resultant acceleration was intended to discover latent patterns in the data that could potentially be useful. This approach had helped to identify landmarks for registration in the power curves in previous chapters (Sections 4.4.2 & 6.4.4). The derivative curves highlighted rates of change, which were thought to be relevant to peak power given that previous studies have shown that RFD is correlated with jump height (De Ruiter et al., 2006; Marcora & Miller, 2000; McLellan et al., 2011). The rate of change in this context was derived from the inertial accelerations at an anatomical location close to the body's CM (LB sensor). On the other hand, studies above often refer to RFD obtained in separate isometric or concentric tests using purpose-built apparatus (De Ruiter et al., 2006; Marcora & Miller, 2000; McErlain-Naylor et al., 2014; McLellan et al., 2011). When discrete variables related to RFD were computed directly from the CMJ VGRF curves, the correlations ranged from 0.138 for the peak RFD to 0.027 for the average RFD (Dowling & Vamos, 1993). Hence, there may be no close correspondence between RFD in those circumstances and the rates of change in inertial acceleration recorded in the current investigation.

Further research indicated that the VGRF's maximum was a more relevant factor than its rate of increase (Tillin et al., 2013). This helps to explain why the ACC components were preferred over rate-of-change (derivative) FPCs. The derivative curve FPCs will have described minor variations in the inertial accelerations that were effectively magnified by taking the derivative. Many studies have shown that RFD has a low jump-to-jump reliability (Cormack, Newton, McGuigan, et al., 2008; McLellan et al.,

2011; Taylor et al., 2010). As a result, including derivative FPCs in the model would have led to overfitting and a higher cross-validated error.

In contrast, the integration curves diminished the effects of these minor variations, and hence such curves contained less information as a whole compared to the original acceleration curves. A few of the FPCs representing cumulative effects proved to be valuable to the model, notably PWR1, VEL2 and PWR2, in that order (Figure 7-9). The pseudo-power components had an intuitive appeal because they attempted to replicate the VGRF-based power curve, as was apparent in the plots (Figure 7-6C & D). However, whilst there was a resemblance, there was not a direct correspondence that correlated highly with the criterion peak power. The limitation of integration-dependent curves is the inherent problem of drift, which cannot be fully addressed by orientation-correction algorithms (Section 2.3.5). In this instance, it was compounded by using the resultant signal, which removed directional information. The FPC plots of pseudo-velocity illustrated this well as the mean velocity continued to increase rather than return to zero (Figure 7-6A & B). However, one of the benefits of FPCA is its ability to extract common characteristics such as drift. VEL1 represented variations in the degree of drift, explaining 75.2% of the curve variance but only 0.2% of the model variance (Figure 7-9). Therefore, VEL2 and VEL4 were independent of drift, making them more valuable to the model. The pseudo-power curves were much less affected by drift as they were products of the acceleration curve as well as the pseudo-velocity. This is evident in the similar levels of explained curve variance for ACC1 and PWR1 (41.2% vs. 45.1%).

The importance of these VEL and PWR components was confirmed by Monte Carlo selection, which relied on GLMSELECT's competitive selection procedure. The procedure weighed up the contributions made by each component to the cross-validated AM loss. In this way, strong predictors were chosen over weaker ones. Hence, the ACC components that were in the majority tended to make a larger reduction to the AM loss than the other components. It does not mean that those components had no relevance, only that they had less relevance to peak power in comparison. Therefore, the multi-curve approach had its merits, but overall such models could not outperform models based on the original acceleration components.

7.4.3 Three-dimensional representations

The inertial accelerations were predominantly in the XZ-plane of the sensor, the sagittal plane in the global reference frame. The accelerations in the X and Z axes were at least an order of magnitude higher than in the Y-axis (mediolateral direction), which were approximately < 0.1 $W \cdot kg^{-1}$. Hence, this would explain why none of the Y-axis FPCs were selected. The relevant characteristic features were almost all in approximately the vertical direction (Figure 7-3). In the sensor's local reference frame, all the X-axis FPCs were chosen along with only two Z-axis components. This reflects the fact that peak power was calculated using only the vertical component of the GRF and strongly indicates that the sensor's changing orientation in the sagittal plane is a minor consideration. Hence, the directionless resultant signal can effectively represent the key characteristics. Moreover, it follows that the 3D signal does not contain much additional information for the model given that the triaxial FPC are correlated between dimensions.

The problem with using the 3D signal was that changes in the sensor's orientation could potentially introduce bias in the inertial accelerations measured along each axis. For instance, a proportion of the acceleration that would otherwise have appeared in the X-axis would appear in the Z-axis, and this proportion would vary throughout the jump. If θ is the deviation angle with respect to some reference vector, such as the initial orientation, then the bias would be proportional to $\sin\theta$. Although initially small, it would increase rapidly with a rate of change proportional to $\cos\theta$. There is no way to correct for this using a sensor with only an accelerometer onboard because its orientation can only be determined when stationary. On the other hand, IMUs with accelerometers, gyroscopes and (optionally) magnetometers can determine sensor orientation with the aid of a fusion algorithm (Luinge & Veltink, 2005; Mazzà et al., 2012; Roetenberg et al., 2007). However, as noted in Chapter 2, such orientation estimates typically involve errors amounting to several degrees on average. The Kalman filters are slow to respond to a change of direction, as occurs in the countermovement jump because they attempt to predict orientation in advance (Figure 2-1). The error may be largest during the propulsion phase when the algorithm responds to the change of direction. These limitations partly motivated this research

into using patterns to estimate peak power rather than trying to obtain accurate measurements directly from body-worn IMU data (Section 2.3.5). Since the jumping movement is strictly controlled (Aragón, 2000; Hatze, 1998; G. Markovic et al., 2004), changes of inclination will modify the inertial accelerations in a consistent fashion. Although the FPCs would describe different shapes as a result, the scores would be essentially the same. Hence, changes of orientation would not be a concern.

The difficulty may lie in the differences in movement strategies between individuals that arise from variations in technique, skill and strength levels. Morphology would also play a role (mass distribution and body segmental lengths), as would fatigue. These factors may be expected to influence sensor orientation as a function of time in response to changes of trunk inclination. Trunk inclination in the sagittal plane is not simply a matter of the hip flexion or extension but also depends on the compound effect through the kinetic chain of knee flexion/extension and ankle dorsi/plantarflexion (Bobbert & van Ingen Schenau, 1988; Feltner et al., 2004; Luhtanen & Komi, 1978). Small changes in lumbar curvature may also have a lesser part to play through the action of the lower back extensors. Finally, the kinematics of how these factors play out for a given individual will vary from one jump to the next.

The models based on the 3D signals were more sensitive to these factors, which may account for the larger errors observed compared to the resultant model. But the difference in predictive error was not large, implying that these factors were not wholly detrimental. The 3D model accounted for the interplay between accelerations in the X- and Z-axes, although the X-axis accelerations were the predominant factor. The 3D model did place emphasis on two Z-axis features, which accounted primarily for the trunk's tilt forwards and then backwards during the countermovement. In conclusion, the inertial accelerations in the X- and Z-axes interacted in a sufficiently consistent way across participants and individual jumps that any such deviations had only a minor effect on the results.

7.4.4 Varimax

The varimax rotation of the FPCs offered the final way to improve the accuracy of the model. It made a marginal improvement in the models based on resultant curves

(0.09 $W \cdot kg^{-1}$), but there was no benefit for the model based on orthogonal curves. Varimax is typically used to aid the interpretation of the FPCs by distributing the curve variance to narrower regions. The question for this investigation was whether, by reshaping the FPCs in this way, some components would describe characteristics that better reflected the peak power output. On the face of it, the correlations for all but one of the 15 varimax-rotated ACC components were higher compared to the corresponding unrotated FPCs. However, the varimax FPCs' correlations partly overlapped, providing no additional information. Hence, the slight improvement observed was well within the noise level, and so varimax was not helpful to the model.

7.4.5 Absence of overfitting

It was notable that as unrotated FPCs were added to the model, the predictive error did not subsequently rise once it had dropped to the lowest value, as would normally be expected. Generally, a model with too many predictors would become attuned to spurious minor variations that, in general, have no relationship with the outcome variable. The GPR model's unexpected behaviour in this respect arose partly because the FPC score magnitudes drop asymptotically towards zero along with the proportion of explained variance. Hence, progressively adding higher-order FPCs would have less and less influence on the model. This situation arose because, in Chapter 5, it had been established that standardisation was not optimal. It can now be understood as a consequence of the FPCs correlation with peak power dropping monotonically (Figure 7-2). Hence, the FPC scores' diminishing size is a factor in the GPR model's apparent lack of overfitting.

The other part of the explanation may lie with the nature of the GPR model itself. The GPR model is based on the Bayesian viewpoint that differs from the conventional, frequentist view, underpinning the LR model. In the Bayesian approach, the parameters are expressed as a distribution of possible values rather than being fixed. The GPR models are fitted based on a range of possible parameter values rather than a range of data sets (Bishop, 2006). Each new observation provides an update to the prior distribution. Bayesian models can accommodate many more parameters or predictors by automatically adapting the effective number of predictors (Bishop,

2006). The SM shows a minimum number of FPCs is required (12), and beyond that, the model can accommodate additional components without increasing the loss (Figure 7-7A). The original assumption in Chapter 5 of 15 FPCs turned out to be close to the optimal number, and so feature selection has had minimal impact on the model's predictive error.

7.4.6 Summary

No further meaningful improvement in the accuracy of the GPR model could be achieved with feature selection. The FPCA-based model benefited little from feature selection because the curve variance explained by the unrotated FPCs dropped off in line with their relative influence on the model. There was no instance of one FPC standing out from that trend, as was the case with the VGRF models. The best GPR model achieved a predictive error of ~2.6 $W \cdot kg^{-1}$, equivalent to 5.8%. This error is larger than the typical intra-day variability of 1.75 $W \cdot kg^{-1}$, the nominal target for the accelerometer model development (Figure 6-10). However, the investigation yielded important insights into the nature of FPCA-based models when applied to accelerometer data, thereby advancing the overall aims of the thesis. This chapter also afforded the opportunity to inspect the nature of the FPCs to understand what characteristic patterns of inertial acceleration were related to peak power output. This is one of the benefits of FPCA as it allows an intuitive understanding of what each component represents. It revealed that the model went beyond a measure of flight time by refining its predictions with many other components. The next chapter will consider the influence of sample size on the model and whether data augmentation techniques may bring about a final improvement in accuracy.

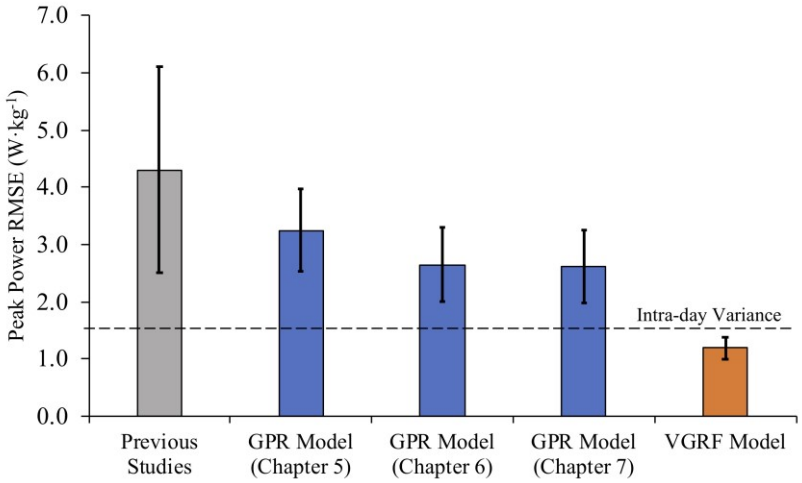


Figure 7-10. GPR models' peak power prediction errors from chapters 5, 6 and 7 in comparison with previous studies (Ache-Dias et al., 2016; Amonette et al., 2012; Canavan & Vescovi, 2004; Lara et al., 2006; Quagliarella et al., 2010; Tessier et al., 2013). MCCV estimate shown for GPR with error bars based on the SD in validation error between folds. The intra-day variance is averaged from four reliability studies (Cormack, Newton, McGuigan, et al., 2008; Hori et al., 2007; McLellan et al., 2011; Taylor et al., 2012).

CHAPTER 8. SAMPLE SIZE AND DATA AUGMENTATION

8.1 Introduction

Over the last three chapters, various aspects of the accelerometer model have been investigated, including algorithm selection, data preprocessing and feature selection. Despite optimisation, the models' predictive error could not be reduced to the same level as the intra-day variability. The models were based on the data set for the CMJ_{NA}, recorded by the LB-attached sensor, which yielded consistently lower error estimates than other sensor/jump-type combinations. Therefore, the final stage of this thesis is to investigate how the model responds to variations in sample size. Hence, this chapter will address the final research question, *how does sample size and composition affect the model's predictive error, and can data augmentation bring worthwhile improvements in accuracy?*

In biomechanics, as in many disciplines, it can be challenging to recruit participants in sufficient numbers. Data collection can be time consuming and resource intensive, running over many weeks or months. Therefore, obtaining estimates for the model error as a function of sample size would help determine whether extending the present study's data collection could be justified. Increasing sample size can also be achieved by the participants each performing more trials. A few biomechanics papers address the number of trials per participant, but those studies were concerned with group comparisons rather than regression models' predictive error (Bates et al., 1992; Dufek et al., 1995; Forrester, 2015).

The alternative course of action would be to use data augmentation techniques to increase the data set's size artificially. As discussed in Chapter 2, augmentation methods were first developed for classification problems to re-balance data sets. (Japkowicz, 2000; Kubat & Matwin, 1997). The SMOTE algorithm (Synthetic Minority Oversampling TEchnique) is a well-known method that generates new data by interpolation, which has been extended to regression models related to ecology and the automotive industry (Chawla et al., 2002; Torgo et al., 2013). The SMOTER

method (with ‘R’ appended to indicate regression) works in feature space, so in the current study it would generate synthetic FPCs scores that define new acceleration curves. It also requires an estimate of the associated peak power for each synthetic curve.

Augmentation can also be based on the original data with a suitable transformation. For instance, the image recognition field uses rotation and reflection to generate new training images (Shorten & Khoshgoftaar, 2019). Similarly, rotating the accelerometer signal would be akin to reorientating the sensor as if it had been attached to the body at a slightly different angle (Section 7.4.2). Actual misalignments of this kind would not matter if the resultant signal were used, but it may offer a way to take advantage of the triaxial data. This technique has not been applied to triaxial sensor data before, but it has the advantage of not requiring peak power estimates for each synthetic point. This approach would be similar to a recent biomechanics study that generated synthetic accelerometer signals from motion capture data (Mundt et al., 2020). Rotating the signal has the advantage of representing what could happen in practice, giving it an intuitive appeal.

Data augmentation should be directed to regions of the outcome variable’s distribution where accuracy needs improvement, in accordance with utility-based regression (Torgo et al., 2015; Torgo & Ribeiro, 2007). Hence, the aim should be to reduce errors in the peak power range where the model is less accurate and more relevant to the end application. In this case, it will be more valuable to focus on the upper range of the peak power distribution as practitioners in elite and professional sport are more likely to be interested in this system.

The aims of this chapter are therefore to:

- Determine how the model’s predictive error responds to changes in sample size based on subsampling the existing data set;
- Assess the effect of participants performing multiple trials in order to inform future data collections; and
- Evaluate two data augmentation methods above to determine whether such methods can be a viable alternative to more extensive data collections.

8.2 Methods

The first aim was addressed by using resampling techniques to determine how well the model would have performed over a range of sample sizes. Estimates can be made of how the model performs for larger samples by extrapolating the observed trend. The second aim was tackled in a similar way by selecting at random only a certain number of jumps from each participant. The signal rotation (SR) and SMOTER methods depended on the same oversampling and under-sampling strategies in which jumps (referred to as cases) were identified for either augmentation or removal. Each synthetic data point was generated according to parameters particular to each method, the details of which are described below. Both methods were assessed initially with exploratory runs of the model to understand the effect of the key parameters before running an optimisation to determine the best settings. Evaluation depended on statistical measures and inspection of the error distributions. All resampling and augmentation methods were carried out by the AM function.

8.2.1 Varying sample size

A new step was added to the AM function to subsample the training/validation set before all other processing. The data set was truncated by participant rather than by trial in the same manner as in the partitioning procedure. A grid search was performed varying the sample size parameter from 1 to 60, the actual number of participants contributing to the training/validation set. The procedure was repeated 100 times with a varying subset of participants to obtain a generalised distribution for each participant total. One hundred iterations were deemed sufficient for each increment in participant number (rather than 1000 iterations in previous chapters) because the standard error is diminished when there is a trend of 60 points. The optimal GPR Chapter 6 model was evaluated using 10 × 10 MCCV for each iteration. A 10-fold design was chosen for the current analysis because the focus was on predictive accuracy rather than model selection. Three different functions (linear, cubic polynomial and exponential) were fitted to the log-transformed data points to find which gave the best fit. The predictive errors were extrapolated to larger samples using these functions to gauge how the model would perform if the sample size were increased above 60. Predictions based

on the surrogate model were not suitable because, beyond the recorded observations, the basis function heavily influenced the Gaussian Process's behaviour.

8.2.2 Varying jumps per participant

The above procedure was then repeated with the additional step of choosing only a random subset of the jumps performed by each participant. The grid search varied the trial-per-participant parameter between 1 and 4. Retaining four jumps per trial was not the same as the original data set because it removed (at random) trials from the 9 participants who had performed eight jumps.

8.2.3 Error distribution fitting method

The next stage of the investigation focused on data augmentation using the full training/validation data set as the starting point – the holdout data set was not included in the calculations. In its current form, the model was more accurate for jumps towards the middle of the peak power distribution where there were more trials (Figure 8-1). The plot shows the absolute error as a percentage to avoid a misleading bias for increasing peak power. The best fit was based on a Gaussian Process rather than a parametric function because a GP has the necessary flexibility and provides a confidence interval (SD). A standard deviation could not be computed with a parametric function at each point because the points were not spaced uniformly along the peak power axis at discrete points. The GP was defined using a pure quadratic basis function and a rational quadratic kernel, which were best suited to the parabolic distribution. The noise level was fitted to data, but this could produce a jagged curve if σ was small. Therefore, the predictive curves were smoothed using a moving average with a Gaussian window of 4 $W \cdot kg^{-1}$. This same fitting method was applied to the augmentation results.

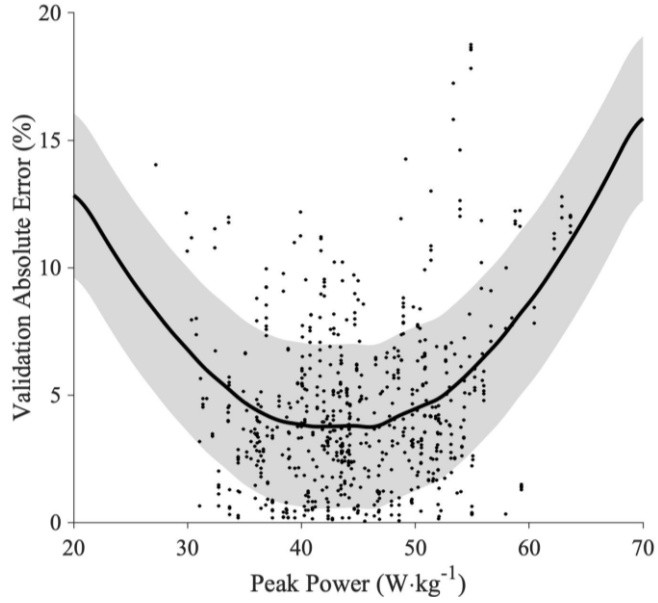


Figure 8-1. Baseline validation error distribution as a function of peak power with no data augmentation for the optimal Chapter 6 GPR model. The solid black line in the central prediction of the Gaussian Process fitted to the data. The shaded area is the GP's confidence interval (SD). Only a quarter of all points are shown (sampled at random) to preserve clarity. Percentage errors shown to remove positive bias for increasing peak power.

8.2.4 Selecting trials for over- and under-sampling

The augmentation strategy was broken down into two parts: oversampling to generate synthetic data and under-sampling to remove some cases from the original set. Research has shown that under-sampling can be more effective than oversampling when predicting values towards the tails of the distribution (Japkowicz, 2000; Kotsiantis et al., 2006; Kubat & Matwin, 1997). Both techniques were performed on the training data alone so that the validation data comprised only real-world data. The following procedures were repeated for each data split of the cross-validation process. The aim was to use a weighted distribution to identify cases for augmentation in an approach similar to ADASYN (Haibo He et al., 2008).

The AM function computed weightings proportional to the outcome variable's inverted probability density function, which was fitted with a kernel method based on a normal distribution. Thus, jumps in the sparse regions of the peak power distribution were more heavily weighted than elsewhere. The function shortlisted a specified number of jumps for oversampling using random sampling with replacement. Jumps

were selected at random with the cumulative weighting distribution, so more heavily weighted jumps were more likely to be chosen (Figure 8-2). The oversampling ratio dictated the number of augmented cases (a multiple of the training set size, [0…5]). The shortlisted jumps could contain duplicates, but this was not a concern because the stochastic procedures used to generate the augmented cases would yield varied results (Sections 8.2.5 & 8.2.6). Probability-based sampling was preferred to a threshold approach (Chawla et al., 2002; Torgo et al., 2015) because it ensured the same number of cases were identified for each CV data split.

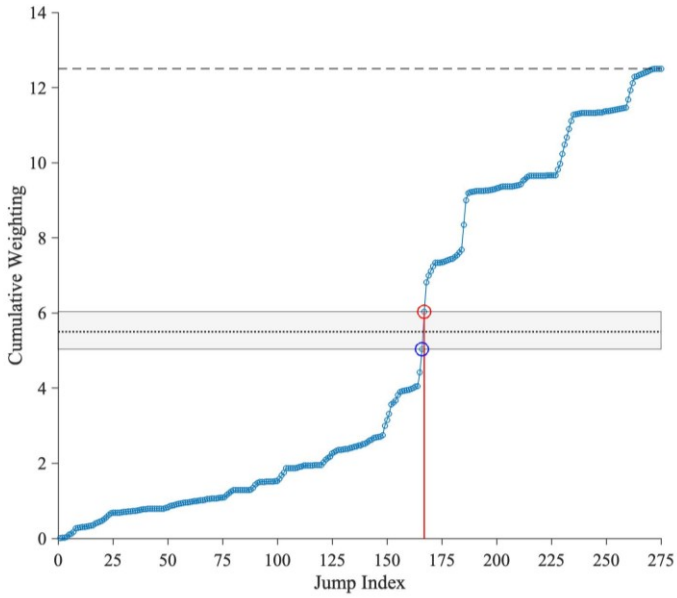


Figure 8-2. Weighted jump selection based on each jump's cumulative weightings across the training/validation set – read across and down. Points are chosen on the vertical axis (dotted line) with a uniform distribution between 0 and the upper limit (dashed line). The case selected (index 167 in this example) is the next point above the dotted line (circled red). Parts of the curve with larger step-ups are more likely to be chosen than smaller increments, as shown with the horizontal band. Consequently, the uniform random selection takes on the cumulative weightings distribution.

A similar approach was taken for selecting jumps for under-sampling. The AM function inverted the weightings and generated a new cumulative distribution. Thus, jumps that previously were associated with large step-ups in the cumulative weightings now made small increments, and vice versa. The algorithm selected cases *without* replacement so that the specified proportion, given by the under-sampling ratio

[0.0…0.9], were removed. The higher the ratio, the bigger the reduction, and like its oversampling counterpart, zero implied no change.

8.2.5 Augmentation – Signal Rotations

The first augmentation technique involved performing random signal rotations (SR) on the raw acceleration data, which formed 3D, time-dependent vectors, $\mathbf{a}_n(t)$. For each jump shortlisted for oversampling, a random angle was generated from a zero-centred Gaussian distribution with a standard deviation specified by a model parameter [0…30°]. A broad range was specified in case the optimal deviation was much larger than expected, although 30° was considered a practical limit. Another parameter defined the rotation axis, and hence the appropriate rotation matrix, **R**, was computed. Each jump's time series was rotated, which involved matrix multiplications, $\mathbf{R}\mathbf{a}_n(t)$ for all *t*. The rotations were performed in either the sensor's local reference frame or the global reference frame. If the latter, an initial time series rotation was performed to reorientate $\mathbf{a}_n(t=0)$ with the vertical axis vector, (-1, 0, 0), using a global rotation matrix, **G**, as described previously (Section 7.2.4). The random rotation was then carried out, followed by the inverse global rotation, reversing the initial change of angle: $\mathbf{G}^{-1}\mathbf{R}\mathbf{G}\mathbf{a}_n(t)$. The rotations about the axes in the global reference frame removed the possibility of individual variations in lumbar curvature affecting the optimisation.

8.2.6 Augmentation – SMOTER

The SMOTER algorithm worked with each shortlisted point in feature space, referred to here as the vector, **u**. First, the algorithm identified the *K* nearest neighbours to **u** across the whole training set, based on the Euclidean distance. The KD Tree method (Friedman et al., 1977) was preferred as it was more efficient on average than an exhaustive search when the number of dimensions (FPCs) was ≤ 8. Above this, the higher cost was not excessive. The algorithm calculated distances based on a truncated number of dimensions (i.e., a set number of FPCs). If distances were measured in dimensions more correlated with the outcome, then the nearest neighbours would be more likely to have a similar peak power to **u**. From the previous chapter, the FPC

correlations with peak power dropped in sequential order (Figure 7–5). The similarity mattered because if the points were close together in terms of outcome, not just in feature space, then linear interpolation was expected to provide a better approximation.

The algorithm randomly chose one of the nearest neighbours, designated as **v**. It defined the augmented data point as, $\mathbf{u}^* = \mathbf{u} + q(\mathbf{v}\text{-}\mathbf{u})$, as a random distance in between **u** and **v**, specified by $q \in [0 \ldots 1]$. It obtained the estimated outcome for $\mathbf{u}^*$ by including the outcome value in **u** and **v** as an extra dimension. This approach was more straightforward giving the same result rather than the formula used by Torgo et al. (2013), who used distance calculations unnecessarily rather than linear algebra. Thus, the augmented data point was defined by linear interpolation between two neighbouring points in feature space, with an estimated outcome determined in the same manner.

Selecting cases based on the outcome distribution would not necessarily correspond to sparse regions of the feature space. Jumps with a high peak power output were less common, but that did not mean the FPC scores necessarily took on extreme values. FPC1 had the strongest correlation with peak power ($r^2 = 0.66$), making this more likely to be the case, but in many instances, it would not. It was even less likely for other FPCs, which had much weaker correlations with peak power. Since the objective was to populate sparse regions of feature space, it was essential to identify points in those areas by updating the weightings calculation to account for the FPC score distributions. Therefore, the overall weighting was computed based on the product of all weightings:

$$\hat{\mathbf{w}} = \prod_{i=0}^{\mathcal{W}} \mathbf{w}_i \tag{8.1}$$

where $\mathbf{w}_i$ is the weighting for the *i*-th FPC, $\mathcal{W}$ is the number of weightings and $\mathbf{w}_0$ is the weighting based on outcome. The weightings for the FPC scores were computed using the same method as for the outcome variable (Section 8.2.4).

8.2.7 Evaluation methods

The parameters required to specify these two data augmentation strategies are summarised in Table 8-1. The augmentation methods were evaluated using 100 × 10 MCCV for a given set of parameters for initial exploratory investigations. Previous chapters had used 1000 iterations, but in this case the number of iterations had to be limited due to the considerable cost in generating the augmented data and fitting the models. The results had a slightly higher variance between runs (± 0.03 $W \cdot kg^{-1}$ vs. ± 0.01$W \cdot kg^{-1}$), but this disparity was deemed minor compared to the scale of the reduction needed (~ 0.8 $W \cdot kg^{-1}$) to achieve a model error equivalent to the intra-day variability.

Table 8-1. Summary of optimisation parameters for data augmentation.

Type	Parameter	Type †	Parameter Values (AM)	PSO Initial Bounds (SM)
Sampling ‡	Over-sampling Ratio	R	0…5	[0, 5]
	Under-sampling Ratio	R	0…0.9	[0, 0.9]
SR	Random Rotation Angle SD	R	0…30°	[0, 30]
	Rotation Axis	C	{X, Y, Z}	[0.50, 3.49]
	Reference Frame	C	{Local, Global}	[0.50, 2.49]
SMOTER	Number of Weightings	I	{0…15}	[-0.49, 15.49]
	Number of Nearest Neighbours	I	{1…10}	[0.50, 10.49]
	Number of FPCs for Distance Calculation	I	{1…15}	[0.50, 15.49]

† Parameter type: C = Categorical (nominal variable); I = Integer; R = Real (continuous).
‡ Optimisation was performed with the Sampling parameters and either SR or SMOTER parameters.

8.2.8 Optimisation

The optimisation procedure was used to determine the optimal values for each parameter in Table 8-1. As it incorporated a random search, it was more efficient than varying one parameter at a time (whilst holding the others constant). Separate optimisations were run for the signal rotations and SMOTER strategies that included the sampling parameters that were common to both. Since the parameters made no changes to the model itself, a nested cross validation design was unnecessary. Instead,

a 1000 × 10 MCCV design was employed in which the SM model was re-trained every 20 iterations, generating 50 models in total. The AM function was modified to return the fourth quartile loss to direct the optimisation to improve predictions for jumps in this top performance bracket using the sampling and augmentation parameters at its disposal.

8.3 Results

8.3.1 Sample Size

The validation loss dropped as the number of participants included was increased, but the rate of decrease slowed (Figure 8-3A). Extrapolating up to 100 participants based on the linear log-transformed fit suggested that even with this larger number of athletes, the model would still fall well short of the target of 1.75 $W \cdot kg^{-1}$ (intra-day variability). A polynomial fit of the log-transformed values achieved the best fit (Figure 8-3B), but the equation suggested approximately 15000 participants would be required (Table 8-2). The exponent function had a similar fit but indicated 740 participants would be needed, a much smaller number. These estimates are uncertain because even a 1% variation in the equation's coefficients could alter the forecast substantially: 650–820 participants for the exponential fit and 9500–25000 participants for the polynomial fit. The fitted equations were more sensitive to the model's predictive errors for a few participants where the points on the graph were more separated.

The predictive errors were lower on average (using the best fit), with each additional jump included per person (Figure 8-4). The largest gain came with the progression from one to two jumps with diminishing returns with each additional jump. The four jumps per participant data set excluded those jumps from some individuals who provided data from eight trials. When those extra jumps were included (equivalent to the data points in Figure 8-3A), there was a further error reduction. With 60 participants, increasing the number of jumps performed from one to four reduced the predictive error by 0.25 $W \cdot kg^{-1}$ (Table 8-3). In comparison, for a fixed sample size of 60 jumps, shifting the balance towards more participants and fewer jumps performed

by each reduced the error by 0.70 W·kg^{-1} (Table 8-4). For example, a more accurate model could be obtained by 60 individuals performing one jump rather than 30 individuals performing two (3.01 W·kg^{-1} vs. 3.26 W·kg^{-1}).

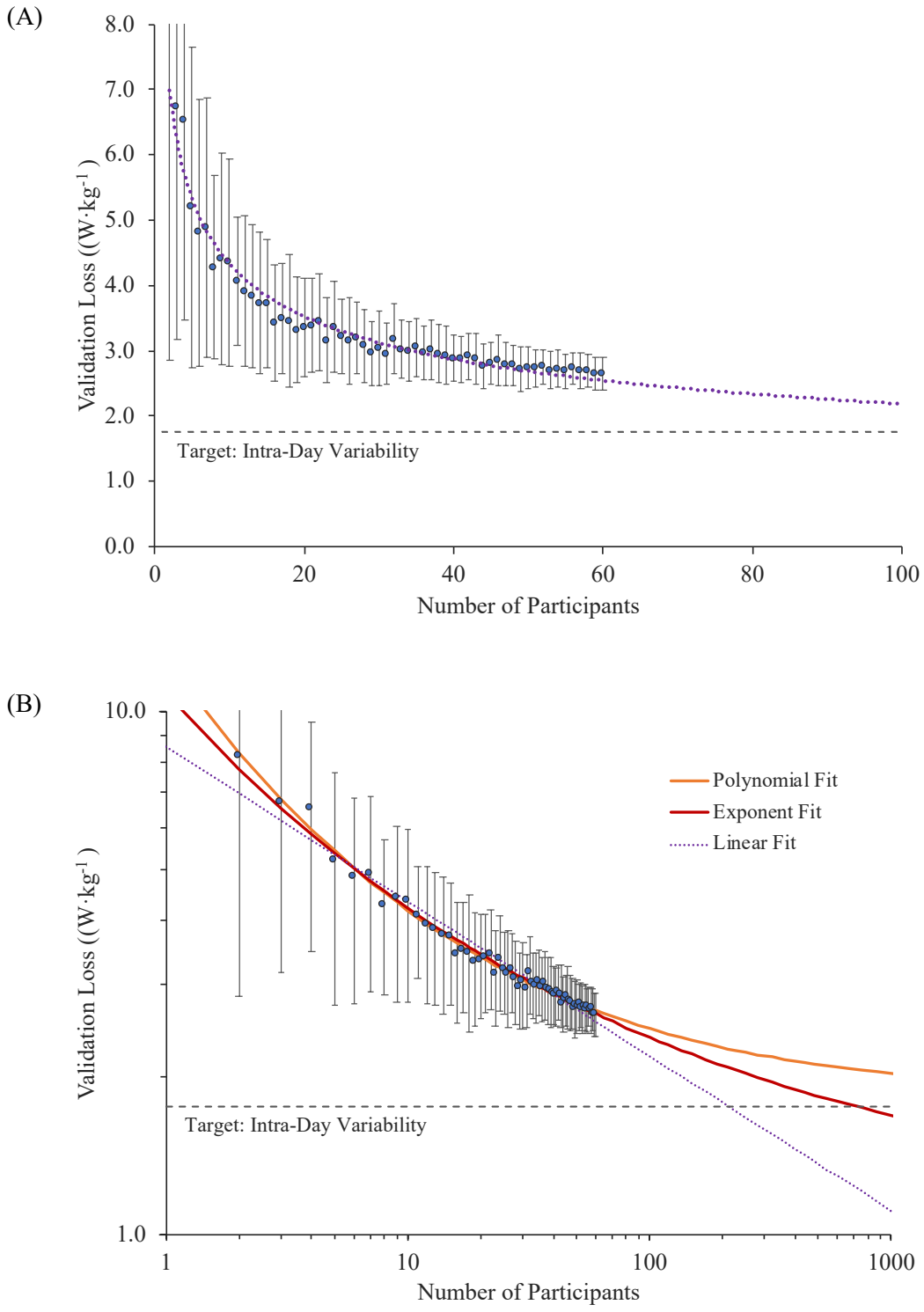


Figure 8-3. Model learning curves: validation loss as a function of sample size with reference to the intra-day variability (1.75 W·kg^{-1}). Data points are mean losses across 100 subsamples, and error bars are their standard deviation. (A) Linear log-transformed fit. (B)

Revised fits of the log-transformed data based on a cubic polynomial, an exponential fit and the original fit from (A). Loss evaluated based on a 10-fold MCCV.

Table 8-2. Best fits for validation loss as a function of the number of participants with forward projections.

	Linear Fit †	Cubic Polynomial Fit †	Natural Exponent Fit †
Fit RMSE (W·kg^{-1})	0.242	0.114	0.143
Fit r^2	95.5%	98.8%	98.5%
Projected Loss for 100 Participants (W·kg^{-1})	2.19	2.47	2.38
Projected Loss for 200 Participants (W·kg^{-1})	1.79	2.28	2.11
Estimated Participants required to meet Target ‡ §	215 [205–225]	15000 [9500–25000]	740 [650–820]

† Type of fit applies to the log-log transformed points shown in Figure 8-3B.
‡ Inter-day variability reported as 1.75 W·kg^{-1}.
§ Top row is the central estimate; bottom row in brackets is the range of estimates from varying the fitted equation's coefficients by ± 1%.

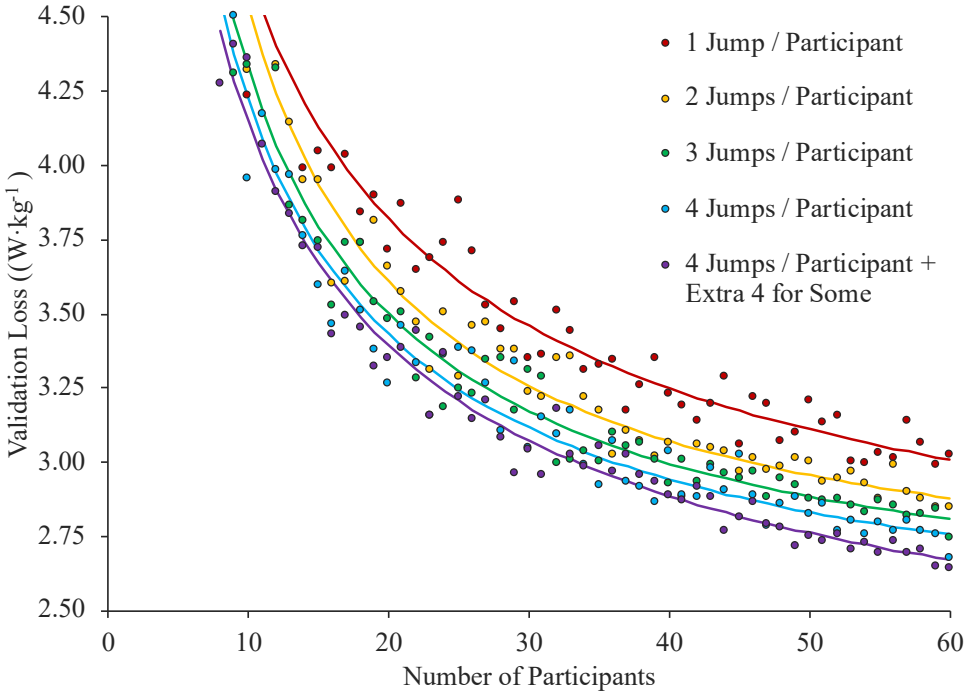


Figure 8-4. Validation loss for a given number of jumps per participant, subsampled at random, as a function of the number of participants, also randomly subsampled. Data points are losses averaged over 100 iterations. Loss evaluated based on 10-fold MCCV. The best fit lines were based on polynomial fit in each case.

Table 8-3. Examples of the estimated validation loss when constrained to 60 participants. Increasing the jumps per participants increases the total sample size.

Constraint	1 Jump / Participant	2 Jumps / Participant	3 Jumps / Participant	4 Jumps / Participant
60 Participants	3.01	2.88	2.81	2.76

Validation loss shown ($W \cdot kg^{-1}$) taken from the best fit of the points shown in Figure 8-4, which were generated by 100 × 10 × 10 MCCV.

Table 8-4. Examples of the estimated validation loss when constrained to 60 jumps, a constant sample size. The balance shifts between participants and jumps.

Constraint	15 Participants / 4 Jumps Each	20 Participants / 3 Jumps Each	30 Participants / 2 Jumps Each	60 Participants / 1 Jump Each
60 Jumps	3.71	3.50	3.26	3.01

Validation loss shown ($W \cdot kg^{-1}$) taken from the best fit of the points shown in Figure 8-4, which were generated by 100 × 10 × 10 MCCV.

8.3.2 Sampling Techniques

The effects of resampling on the training set's peak power distribution are shown in Figure 8-5, along with the corresponding distribution of validation error. The baseline peak power distribution had two peaks (Figure 8-5A, red line). When the oversampling ratio was introduced, those peaks were diminished as the sampling ratio was increased. The curve with an oversampling ratio of 5 had a flatter distribution from the 10^{th} to the 90^{th} percentile, approximately 36–54 $W \cdot kg^{-1}$ (Figure 8-5A, purple line). The validation error distribution had a shape characteristic of the normal distribution when absolute values are taken (Figure 8-5B). Its peak shifted from 1.4 $W \cdot kg^{-1}$ to 0.8 $W \cdot kg^{-1}$ with oversampling.

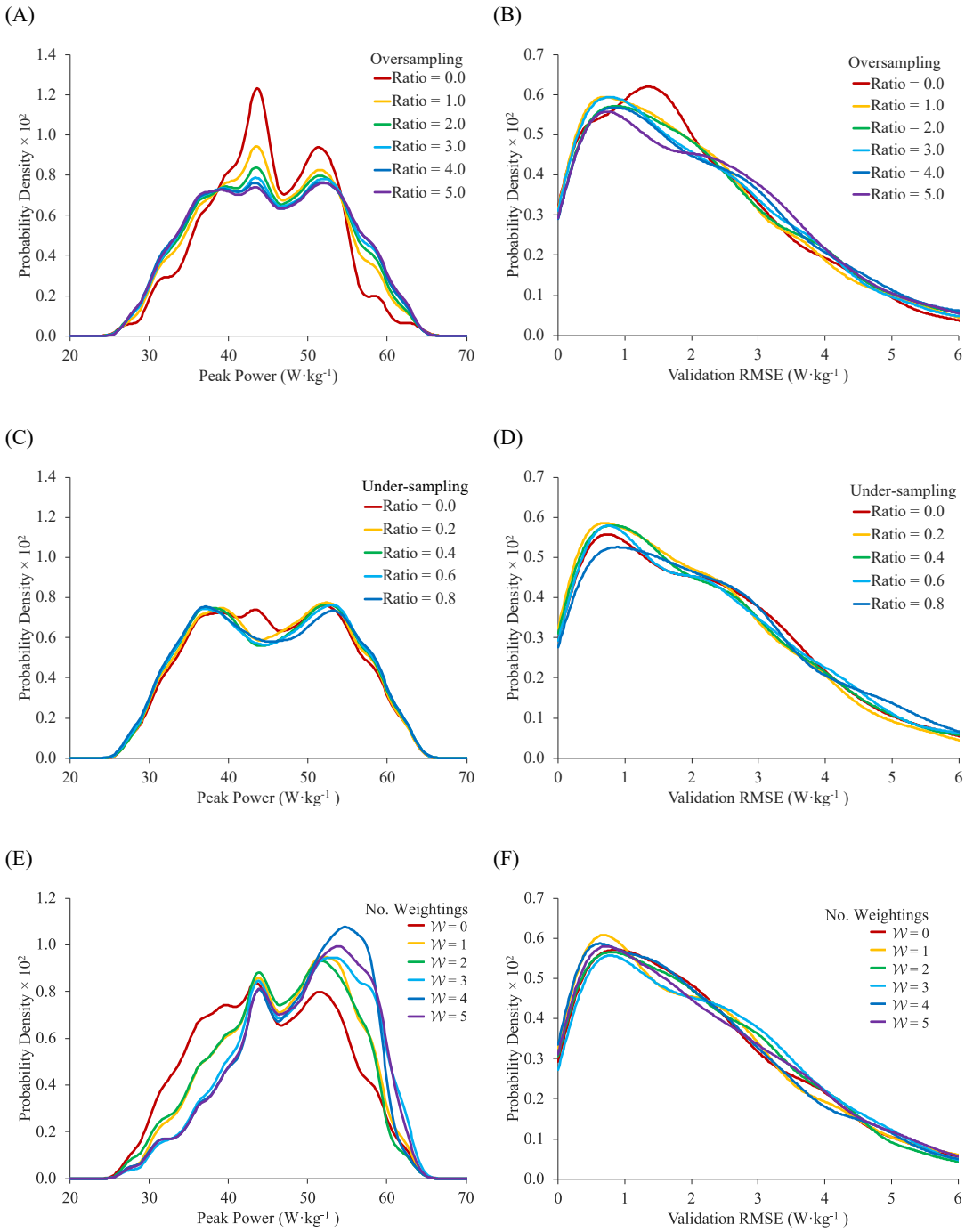


Figure 8-5. Training data distribution following resampling (left-hand column) and the corresponding validation error distribution (right-hand column). Top row: variations in the oversampling when the under-sampling ratio = 0. Middle row: variations in under-sampling when the oversampling ratio = 5. Bottom row: variations in the number of weightings applied when the oversampling ratio = 2 and the under-sampling ratio = 0. Number of weightings = 0 for the top and middle rows. Results shown when running SMOTER augmentation, but similar distributions arise for SR augmentation. Note: small disparities in the distributions may arise when re-running the augmentation procedure due to its stochastic nature.

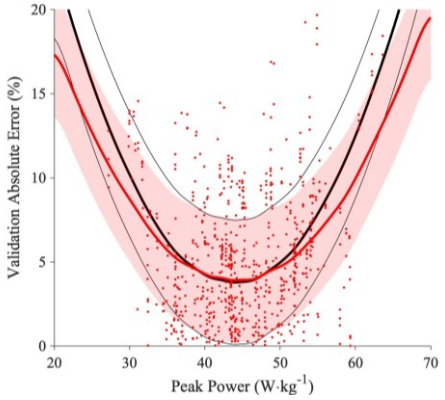
In the second row of Figure 8-5, the baseline distribution was the over-sampled curve (ratio = 5) from a separate run of the augmentation procedure. It differs slightly from the equivalent distribution in the top row due to the stochastic nature of oversampling. Introducing under-sampling induced the probability density to fall in the middle of the range as the under-sampling ratio was increased, leading to peaks either side around 38 $W \cdot kg^{-1}$ and 52 $W \cdot kg^{-1}$ (Figure 8-5C). The validation error's distribution rose a small amount at the peak but was accompanied by slight reductions elsewhere (Figure 8-5D).

On Figure 8-5's third row, from an additional run, the baseline distribution was the same as above with an over-sampling ratio of 5, but with no under-sampling. When weightings were introduced to factor in the FPC scores, the distribution's second peak tended to become more prominent in the 50–60 $W \cdot kg^{-1}$ range (Figure 8-5E). However, increasing $\mathcal{W}$ did not always lead to a higher second peak. Including FPC1, FPC3 and FPC4 in the weighting produced rises, but taking account of FPC2 and FPC5 led to reductions. There were no discernible changes to the validation error distribution when applying differing numbers of weightings (Figure 8-5F).

8.3.3 Validation error vs. peak power

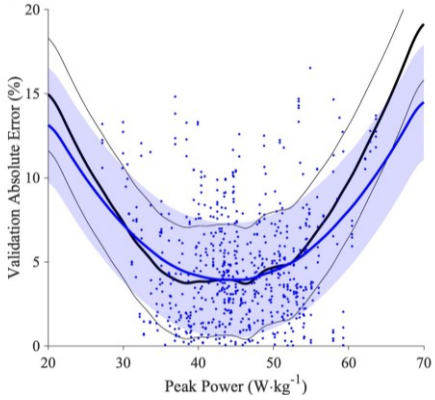
A side-by-side comparison of the two augmentation methods for different resampling strategies is shown in Figure 8-6. Overlaid within each plot is the outline of the baseline case without augmentation, which differs slightly between the 3D signal for SR and the resultant signal for SMOTER. Oversampling reduced errors at the peripheries for both methods (Figure 8-6A–B). Introducing undersampling brought a marked improvement for SMOTER and to a lesser extent for SR (Figure 8-6C–D). Increasing the random rotation angle's SD from 10° to 20° was detrimental to SR (Figure 8-6E). Increasing the number of weightings from 0 to 1 was beneficial for SMOTER (Figure 8-6F). These exemplar plots demonstrate how changes to certain parameters influence the error distribution but finding the best values for all six parameters required optimisation.

(A)



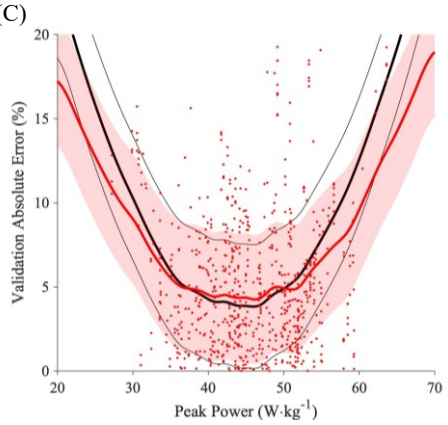
Oversampling Ratio = 2.0; Undersampling Ratio = 0.0; Rotation Angle SD = 10°. Rotation Axis = X (Local)

(B)



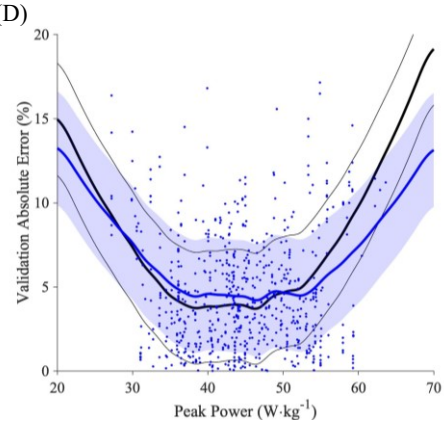
Oversampling Ratio = 2.0; Undersampling Ratio = 0.0; Weightings = 0; NN = 5; FPCs for NN Distance = 15.

(C)



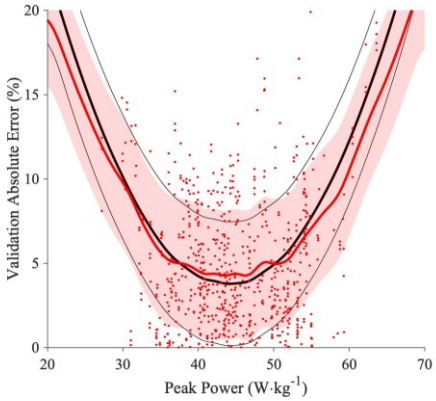
Oversampling Ratio = 2.0; Undersampling Ratio = 0.4; Rotation Angle SD = 10°. Rotation Axis = X (Local)

(D)



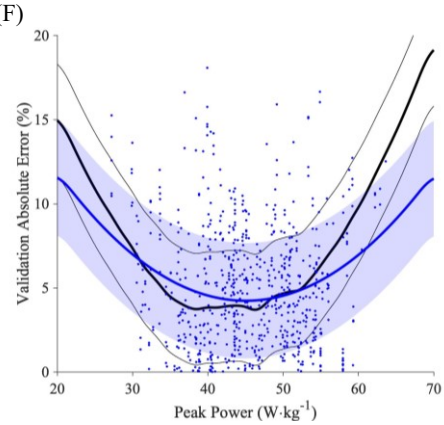
Oversampling Ratio = 2.0; Undersampling Ratio = 0.4; Weightings = 0; NN = 5; FPCs for NN Distance = 15.

(E)



Oversampling Ratio = 2.0; Undersampling Ratio = 0.4; Rotation Angle SD = 20°. Rotation Axis = X (Local)

(F)



Oversampling Ratio = 2.0; Undersampling Ratio = 0.4; Weightings = 1; NN = 5; FPCs for NN Distance = 15.

Figure 8-6. Examples of over- and under-sampling on the distribution of absolute validation errors across the performance range based on SR using 3D signals (left-hand column) and SMOTER using resultant signals (right-hand column). Black lines show the baseline case without augmentation. The shaded area is SD. Top row: oversampling alone. Middle row: oversampling with under-sampling introduced. Bottom row: Over and under-sampling with either wider rotations or more weightings. Percentage errors shown to remove positive bias for increasing peak power. Only a quarter of all points are shown (sampled at random) to preserve clarity. The baseline fits differ between the SMOTER and SR plots mainly due to differences between the resultant and 3D models.

8.3.4 Augmentation optimisation

The SM predicted loss and its associated confidence interval converged after 200 out of 1000 iterations for SR and 360/1000 for SMOTER. However, there was a considerable level of noise (0.65 $W \cdot kg^{-1}$ and 0.50 $W \cdot kg^{-1}$ for SR and SMOTER, respectively), which was much higher than in previous optimisations (0.2–0.4 $W \cdot kg^{-1}$). The parameters for both methods did not show signs of stabilisation as the optimisation progressed. Nevertheless, the optimal parameter settings could be determined using the modal definition (Table 8-5). Both methods performed best with a high level of oversampling, 4.65–4.70 times the original sample, such that only around 17% of the training data was from the data collection. The SR method preferred removing about two-thirds of this data, whilst SMOTER kept almost all the original data. The optimal standard deviation for the random signal rotations was around 12°, which were best performed about the X-axis in the sensor's reference frame. The optimal SMOTER configuration selected a point from the eight nearest neighbours and based their distances on six dimensions (FPCs). It preferred to base the weightings on the FPC1 and FPC2 scores as well as peak power.

Table 8-5. Optimal, modal-defined parameters for both augmentation strategies based on the Chapter 6 GPR model.

Parameters / Loss	SR	SMOTER
Over-sampling Ratio	4.70	4.65
Under-sampling Ratio	0.68	0.05
Rotation Angle SD (°)	11.8	n/a
Rotation Axis	X	n/a
Reference Frame	Local	n/a
Number of Weightings	n/a	2
Nearest Neighbours	n/a	8
Number of FPCs	n/a	6

Basis Function = None; Kernel Function = Exponential; $\log_{10}\sigma$ = -2.20; Standardisation = No. Alignment = Take-off; t_{pre} = 2000 ms; t_{post} = 1200 ms; $\log_{10}\lambda_{FS}$ = 4.80; No. Basis Functions = 100; Varimax = No.

Using the parameter values above, SMOTER made a more substantial change to the error distribution than the SR (Figure 8-7). It yielded a notable reduction in curvature, which was more beneficial towards the upper range of jump performance.

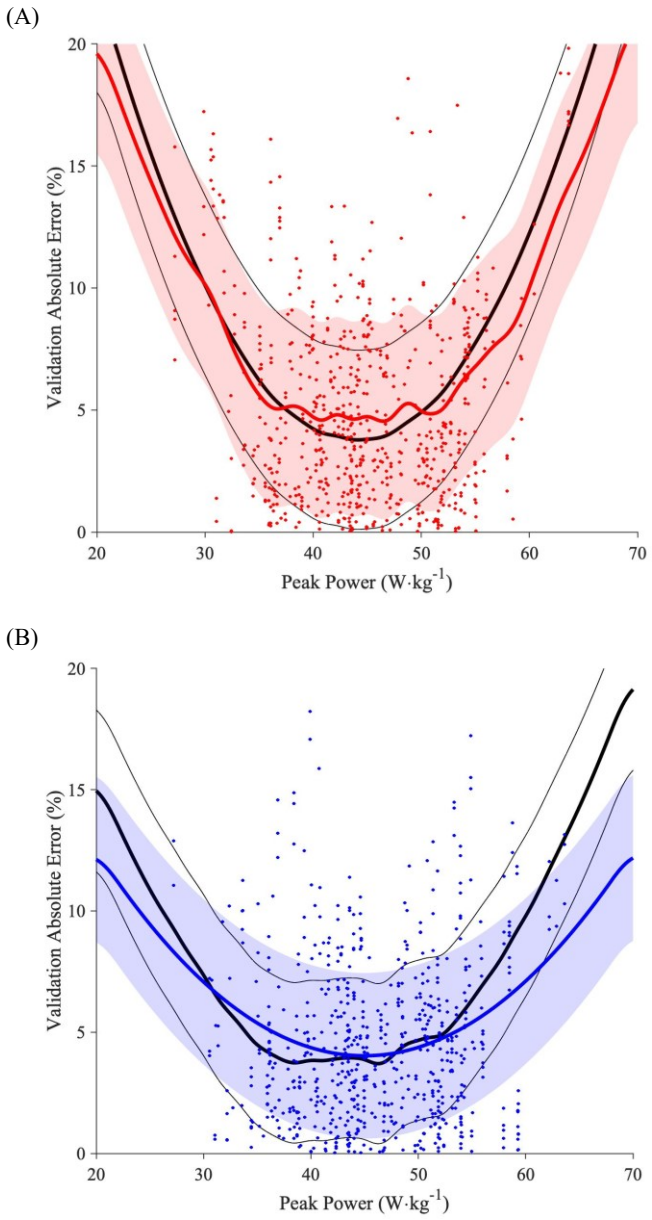


Figure 8-7. Predictive error distribution as a function of peak power for the optimised augmentation methods: (A) SR; (B) SMOTER. Black lines show the baseline case without augmentation. The confidence interval (shaded area) is SD. Percentage errors shown to remove positive bias for increasing peak power. Only a quarter of all points are shown (sampled at random) to preserve clarity. See Table 8-5 for parameter values.

The MCCV loss estimates summarise the changes in the respective error distributions (Table 8-6). Neither method improved the GPR model's overall accuracy, but they did reduce the predictive error for the 4th quartile, as directed in the optimisation. The improvement, however, was only moderate, from 4.28 $W \cdot kg^{-1}$ to 4.00 $W \cdot kg^{-1}$ for SR and from 3.71 $W \cdot kg^{-1}$ to 3.41$W \cdot kg^{-1}$ for SMOTER. A comparison between both methods was possible since SMOTER could be applied to the 3D signal as well, although it had not been optimised for this input. The comparison revealed that SMOTER was more effective at reducing the 4th quartile loss than SR (3.21 $W \cdot kg^{-1}$ vs. 4.00 $W \cdot kg^{-1}$). However, both methods achieved a similar result overall.

Table 8-6. Optimised model performance for both augmentation strategies showing the breakdown of validation by peak power quartile. Comparisons are made with the baseline without augmentation. The SMOTER results are split into those based on the resultant signal (1D) and the 3D signal to allow a like-for-like comparison with SR. Optimisation was based on the Q4 loss.

MCCV Loss ($W \cdot kg^{-1}$) †	Q1	Q2	Q3	Q4	Overall
SR					
Baseline (3D)	2.72	2.84	2.66	4.28	**3.20**
Augmentation (3D)	2.62	2.46	3.20	4.00	**3.16**
SMOTER					
Baseline (1D)	2.14	2.20	2.60	3.71	**2.63**
SMOTER (1D)	2.44	2.66	2.69	3.41	**2.92**
SMOTER (3D)	2.76	2.62	2.63	3.21	**3.21**

† Cross validation design: 100 x 10 MCCV.
With 100 iterations, results can vary between runs by ~ ± 0.03. 1000 iterations would be expected to reduce this variability to ~ ± 0.01 with sufficient computational resources available.
See model and augmentation specification in Table 8-5.

The strength of these optimal values can be gauged by inspecting the parameter distributions and the SM plots, bearing in mind the concerns reported above with the optimisation. For SR, an oversampling ratio around 2 was an alternative with a smaller angular spread when the Y-axis was chosen (Figure 8-8F, H & I, noting the cluster distributions). The SM plot for the random rotation SD exhibits a corresponding minimum along with the undersampling ratio (Figure 8-8B & C).

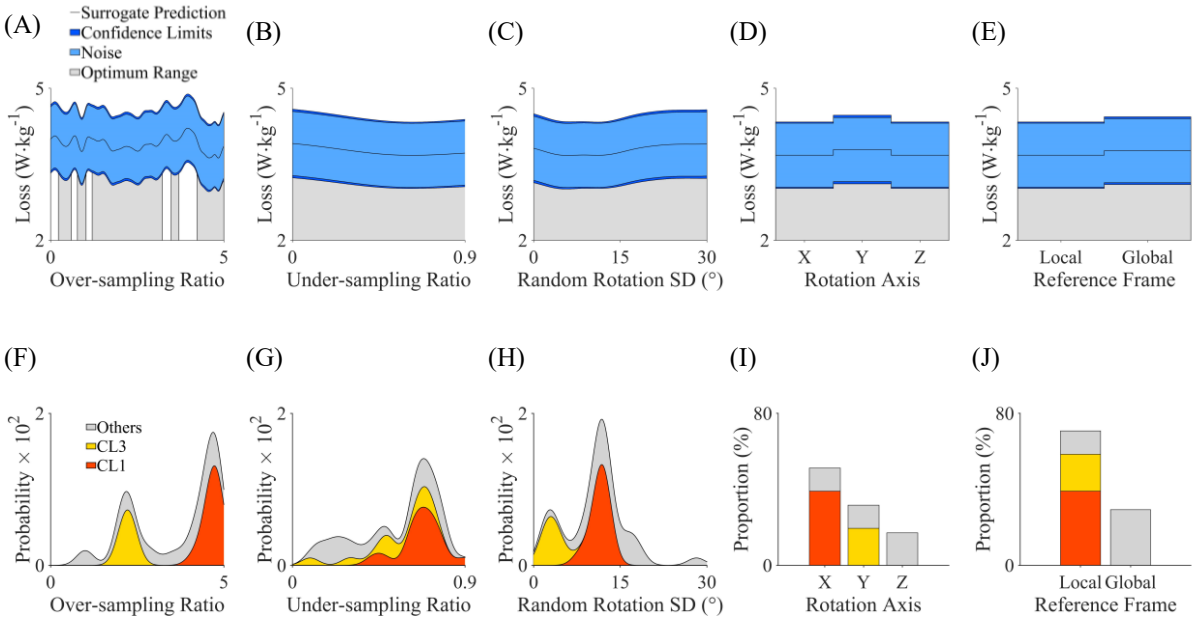


Figure 8-8. Surrogate model plots (top row) and the corresponding parameter distributions (bottom row) for the SR augmentation method. Two clusters (CL1 & CL3) are highlighted.

For SMOTER, the SM was flat around the optimum except for the under-sampling ratio, which showed a drop close to zero on the horizontal axis (Figure 8-9A–E). The corresponding parameter distributions revealed more structure (Figure 8-9F–J), but no clusters could be identified as the parameter values were spread out with no discernible groupings. Only under-sampling showed an unambiguous peak. The oversampling ratio had four peaks with a tendency towards a higher ratio (Figure 8-9F). There were wide spreads in the numbers of nearest neighbours and how many FPCs to use when calculating the distances between them (Figure 8-9I–J). A similar outcome is apparent for the number of weightings, but notably, zero and one were never chosen (Figure 8-9H). $\mathcal{W} = 0$ indicates weightings based solely on peak power.

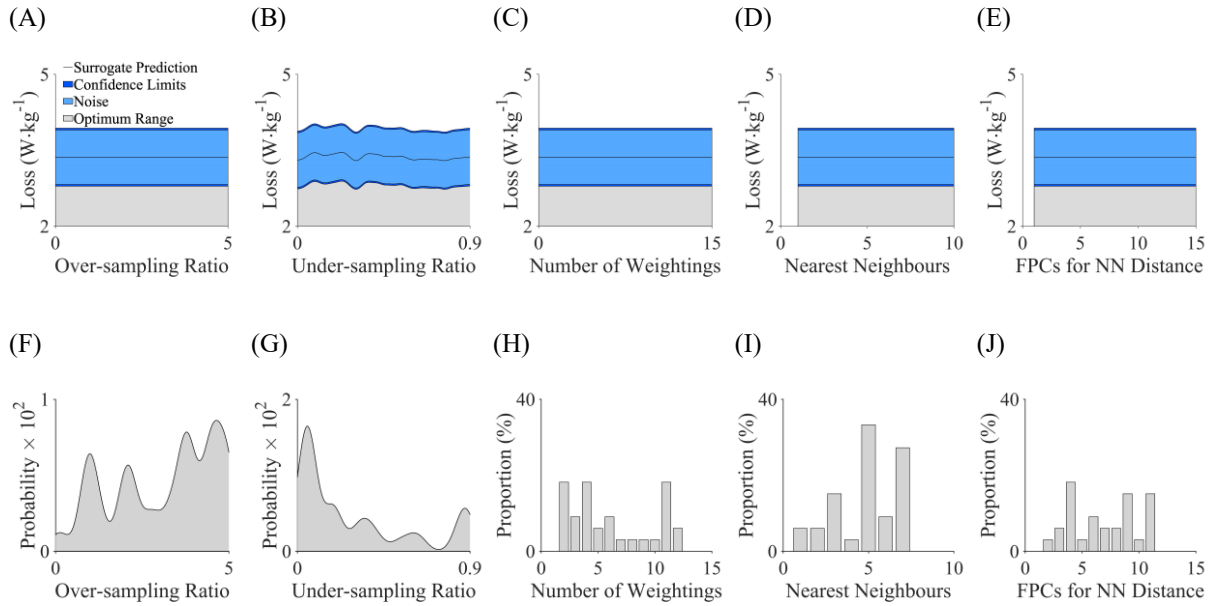


Figure 8-9. Surrogate model plots (top row) and the corresponding parameter distributions (bottom row) for the SMOTER augmentation strategy. The distributions are all grey because no clusters were found. NN = Nearest Neighbour.

8.4 Discussion

This chapter considered the effect on the model's accuracy by varying the data set's size or altering its distribution. The aims were to determine how the predictive error responds to changes in sample size, assess the effect of participants performing multiple jumps, and evaluate whether data augmentation methods can be a viable alternative to larger data collections.

8.4.1 Sample size

The first part of the analysis investigated whether increasing the sample size would reduce the error to the level of intra-day variability (1.75 W·kg^{-1}), and if so, how many participants would be required. It was evident that predictive error had a logarithmic relationship with sample size, rather than being inversely proportional to $\sqrt{N}$, as was presumed when computing the benchmark error for the peak predictive equations (Section 2.3.2). Had those studies' errors been weighted by the $\log N$ instead, the benchmark would have been little different (4.4 W·kg^{-1} vs. 4.6 W·kg^{-1}).

Three projections were made based on different functions fitted to the log-transformed data. By inspection, it was clear that the linear function, which gave the most optimistic projections, was not well fitted to the data. The closest fit was made by the cubic polynomial, which indicated that further error reductions would be quite limited. It projected an error of 2.47 $W \cdot kg^{-1}$ for 100 participants and 2.28 $W \cdot kg^{-1}$ for 200, assuming that the same trend over the first 60 participants continues. Since the polynomial fit had the slowest convergence of the three equations, it produced a very large estimate for the number of participants required to reach the 1.75 $W \cdot kg^{-1}$ target (~15000). The estimate was quite uncertain because of its sensitivity to the equations' coefficients, which were more strongly influenced by the model's predictive errors for small samples (Figure 8-3). Therefore, the exponential function was fitted instead as it had a faster convergence rate. Its fit was almost as good as for the polynomial (r^2 = 98.5% vs. 98.8%), estimating around 740 participants would be required to meet the target. The exponential fit was considered a reasonable compromise between the apparently optimistic and pessimistic estimates from the other two equations.

These projections are inherently uncertain because the extrapolation range is considerable, going far beyond the actual observations of up to 60 participants. Furthermore, in practice additional participants may not fit into the current distribution assumed in the projections. Hence, there is considerable uncertainty over the exact numbers required. However, what is clear is that it would be a very large number of participants. Typically, the largest biomechanics studies rarely approach 100 participants, let alone several hundred or several thousand. Moreover, recruiting large numbers of moderate to high performers is more challenging because, by definition, there are fewer available in the specific populations of interest. When the target sample size is large, a compromise is often unavoidable as it will be more practical to recruit participants with average ability to make up the numbers. As a result, the distribution will shift away from the proficient jumpers, pushing the target participant number even higher. Dedicating time and effort to undertaking such a large data collection would often be hard to justify, given the demands of high-performance environments or on laboratory resources for other priorities. Therefore, whilst modest improvements can be expected from recruiting more participants, on a par with the gains seen in previous

chapters, reducing the model's predictive error to the target of 1.75 $W \cdot kg^{-1}$ is practically out of reach.

However, in one other respect, the analysis served to show that the actual number of participants recruited was large enough to obtain a reasonable estimate of the model's true capability. The lowest error of 2.5 $W \cdot kg^{-1}$ from previous chapters was not much higher than what the projections indicated was ultimately possible with these FPCA models, based on accelerometer data. The rate of improvement had already slowed substantially, limiting what further could be achieved. Hence, the analysis confirmed the data collection was sufficient to fulfil the aims of this thesis.

8.4.2 Trials per athlete

The second aim was to determine the benefit of having participants perform multiple jumps. The results showed that the model's predictive error decreased with every additional jump performed. The biggest gain came from having two trials rather than one, with a law of diminishing returns thereafter. The vast majority of participants performed four jumps of the same type, which helped improve model estimates from 3.01 $W \cdot kg^{-1}$ to 2.75 $W \cdot kg^{-1}$ for the full training/validation data set. The error was lowered to 2.67 $W \cdot kg^{-1}$ by inviting back some participants who had taken part in the first round of data collection, demonstrating that it is worthwhile getting the maximum number of jumps, even if they come from the same individuals. However, given a choice between more participants or more trials, the results showed for the same sample size, the model was more accurate when the balance shifted towards more participants (Table 8-4). Moreover, the error reduction was greater when increasing the number of participants than increasing the number of jumps per participant (0.70 $W \cdot kg^{-1}$ vs. 0.25 $W \cdot kg^{-1}$ from Table 8-3). Hence, the model appeared to learn more from each trial when performed by another individual rather than by the same person. Therefore, every effort should be made to recruit as many participants as possible, but getting the existing recruits to perform extra trials will also yield modest benefits.

Other investigators have established that recruiting more participants increases the power of *t*-tests compared to increasing the number of trials (Bates et al., 1992;

Forrester, 2015). They reported that the required statistical power could be achieved by varying the balance between participants and trials, e.g. a statistical power of 0.80 can be obtained with 15 participants and 19 trials, or 20+ participants and three trials (Forrester, 2015). Those investigations focused on hypothesis testing with straightforward statistical measures (SEM – Standard Error of the Mean), in which analytic expressions can be derived. However, the present study's approach resembles the random sampling method used by Dufek et al. (1995), who simulated different numbers of participants and trials from their running VGRF data in a comparative study. The resampling method is free of assumptions about the model or the data distribution (Davison & Hinkley, 1997) unlike conventional methods, which make several assumptions including linearity, homoscedasticity, error independence and error normality (Atkinson & Nevill, 1998).

Resampling ensured that the jumps performed by any one participant were assigned either to training or validation, but not both. The partitioning of data by participants was implemented in code rather than Matlab's Regression Learner app. If the data were not partitioned in this way, the model would have had a substantial advantage, resulting in over-optimistic lower predictive errors. If a model trained in that way were applied to unseen data, it would perform poorly (Appendix E.2). Hence, this example demonstrates the similarity in characteristics (FPCs) between jumps performed by the same individual. Conversely, it also shows how much more different those characteristics are between participants. Thus, the model benefits more from additional participants rather than additional trials.

These findings suggest that the model benefits from a more heterogenous training data set. However, the error distribution across the peak power range showed that errors were lower when there were more jumps at the same performance level. Therefore, ideally more participants should be recruited in the performance range of interest in the target population, reflecting the specificity principle. This could be taken to mean that participants should come from the same sport as well, but there may be more variability in movement strategies between individuals than across sports. Techniques will vary between individuals reflecting differing proficiency levels, strength and coordination, and anthropometric differences, including the distribution of mass,

segmental lengths, and so on. The sport's demands will have a bearing on some of these factors, or some of these factors may influence the participant's choice of sport. The sport-specific training may lead to similar movement strategies within a given sport, but as jumping is a fundamental skill, there may be some overlap across sports. Hence, there may be advantages to recruiting participants from several related sports as a means to increase the sample size.

8.4.3 Augmentation strategies

The final part of the investigation evaluated two augmentation techniques as an alternative approach to recruiting more participants or increasing the number of jumps performed by each person. In line with utility-based regression and ADASYN, the augmentation was targeted towards the peak power distribution tails where errors were higher. Specifically, the upper end of the performance range was targeted, given the potential application to professional and elite sport. High performers are fewer in number in a population of active sportspeople, by definition, such that it would be challenging to recruit them in sufficient numbers. Based on the evidence from other domains (Section 2.6.4), it was proposed that augmentation may alleviate that problem with synthetic data.

The results showed that the SMOTER and SR methods had the desired effect, reducing the predictive errors in this region. SMOTER achieved a more marked reduction in error towards the distribution tails, which could be seen most clearly in the altered error distributions (Figure 8-7) than in the breakdown of the losses by quartile (Table 8-6). SMOTER also started from a lower baseline as it worked with the resultant signal, which yielded more accurate models, as the previous chapter showed. Although the improvement was not significant, this investigation showed for the first time that, in principle, data augmentation could be applied to accelerometer signals for a modest reduction in error. Nevertheless, augmentation could not fully compensate for the lack of data in the performance range of interest. A model trained on data predominantly from those with moderate jumping abilities, boosted with augmentation at the high end of performance, still produced larger errors when applied to high performers. These high performers may well exhibit somewhat different movement patterns, quantified

by the FPC scores, as a result of superior jumping skill, greater strength or other possible factors. Although heterogeneity is valuable, the model is not necessarily able to take the learned relationships and apply them to a wider grouping. In conclusion, it seems clear that there is no getting around the fact that if high accuracy is needed for a particular cohort, then models must be trained on participants with those abilities.

8.4.4 Manipulating the peak power distribution

This is the first time these augmentation methods have been applied to accelerometer data. Although they did not make a significant difference, the techniques may have applicability elsewhere in biomechanics. It is therefore worth reviewing the two methods in more detail. The purpose of resampling was to shift the model's focus away from frequently occurring observations towards regions of performance that were rare but of more relevance to the end application (Haibo He et al., 2008; Torgo et al., 2013). The over- and under-sampling methods employed were common to both SR and SMOTER as they determined which cases were shortlisted for augmentation or removal. With the correct sampling ratios, the peak power distribution was transformed as desired into an approximately uniform distribution over 90% of its range. (The kernel-based distribution fit was more sensitive to the data than the original histogram, Figure 3-3A). The probability-based selection criterion was more effective at achieving this than the threshold-based approach developed originally for SMOTE (Chawla et al., 2002).

The outcome distribution on its own could determine where augmentation should be directed in the peak power range. However, it was not a good proxy for the predictor distributions, as most FPCs were not well correlated. Moreover, proximity in feature space did not necessarily reflect close similarity between acceleration curves because the first FPC was the biggest factor with subsequent components in descending order of importance. Adjusting the weightings to account for FPC distributions was intended to mitigate this distortion and ensure that the synthetic data populated sparse space regions of feature space. It had a clear impact on the resulting peak power distribution, more so than altering the resampling ratios. The more variables included, the narrower the probability distribution function became, effectively reducing the pool of available

jumps for augmentation. Notably, the probability density rose toward the upper end of the distribution without the algorithm specifically targeting this region (Figure 8-5E). It did so because there were more FPC1 outliers at the top rather than the bottom end of the scale. The exceptional jumpers differentiated themselves more from the other participants than the poor performers did at the other end of the scale. That situation arose by chance in the present study, but it could be enforced by zeroing the weightings below the median if need be. It is believed to be the first time that predictor distributions have been factored into resampling procedures.

8.4.5 Estimating synthetic outcome values

Estimating a value for the outcome variable for a new synthetic point is a fundamental problem for data augmentation in regression models (Torgo et al., 2015) The problem exists for classification models too, but not nearly to the same extent because outcomes are restricted to two or a handful of possibilities. The SMOTER algorithm used interpolation to estimate the outcome variable from the corresponding values from two real data points. However, making a prediction based on two data points is prone to error. One possibility is to increase the number of points on which the estimate is made and use cubic interpolation or other similar techniques. Ultimately, a predictive model of some kind would be required – linear interpolation was the simplest possible model. Such a model trained on a handful of nearest neighbours would not be any more accurate than the full model trained on all the points. However, the purpose of augmentation was not to pre-empt the full model but to shift the model's emphasis towards the outcome range that was more relevant to the practitioner (Torgo et al., 2013). In that sense, the accuracy of synthetic data's outcome variable was of secondary importance. The error distribution change demonstrated augmentation had fulfilled its purpose to the extent that errors rose slightly in the mid-range but fell towards the peripheries (Figure 8-6). Yet, without accurate outcome estimation, the improvements it can bring in the region interest will always be limited.

The problem does not arise for SR, however, because the outcome is assumed to be the same. Its rationale is that different accelerometer signals would have been produced for the same jump had the sensor been orientated slightly differently, a

situation that would inevitably occur in practice. By analogy with image recognition systems, an image is the same at whichever orientation it is viewed (Shorten & Khoshgoftaar, 2019). Alternatively, the augmented data can be derived from another data source, such as using motion capture data (Mundt et al., 2020). Despite its apparent advantages, SR did not perform as well as SMOTER, judged by the degree of change in the error distribution. It also started from a higher baseline error because it must work with 3D signals. SR did not reduce the 4th quartile loss as much as SMOTER, but it did achieve a slight reduction in the 1st quartile loss, something that SMOTER failed to do. SR did not benefit from adjusting the weightings, so it worked at both ends of the peak power distribution.

The SR method's poor performance can be attributed to the profound effect the rotations have on the acceleration curves. Variances in one dimension are transferred into another, substantially changing the curves' shapes. Rotations about the X-axis were favoured because transfers between the Y- and Z-axes were comparatively minor. The random signal orientations only appeared to worsen the accuracy rather than providing the model with a richer set of training examples. In image recognition, rotations typically only yield a marginal improvement in accuracy by a few percentage points (Shorten & Khoshgoftaar, 2019). Crucially, though, the DNNs set about finding patterns in the images in an entirely different way to FPCA, which is limited by the need for close curve alignment.

8.4.6 Summary

The resampling analysis established that a larger data collection could improve the model's accuracy, but only to a limited extent. Only with considerable resources and a large pool of willing participants, possibly exceeding a thousand, might it be possible to lower the current model's predictive error below the target of 1.75 W·kg^{-1}, the intra-day variance (Figure 8-10). There is no guarantee that this would be achieved as the projection was from a comparatively small base. The analysis also confirmed that the training/validation data set used in this thesis was large enough to allow the models to perform close to their full potential.

Increasing the number of jumps per participant yielded a marginal improvement, but what mattered more was the total number of athletes. The results showed that the model learned more from each additional individual than they did from each additional trial by the same person. Data augmentation could only offer a slight improvement in the model's accuracy at the top end of jump performance, where it can be hard to recruit many such athletes. However, a modest reduction in error in this peak power range is far from certain as the errors remained large.

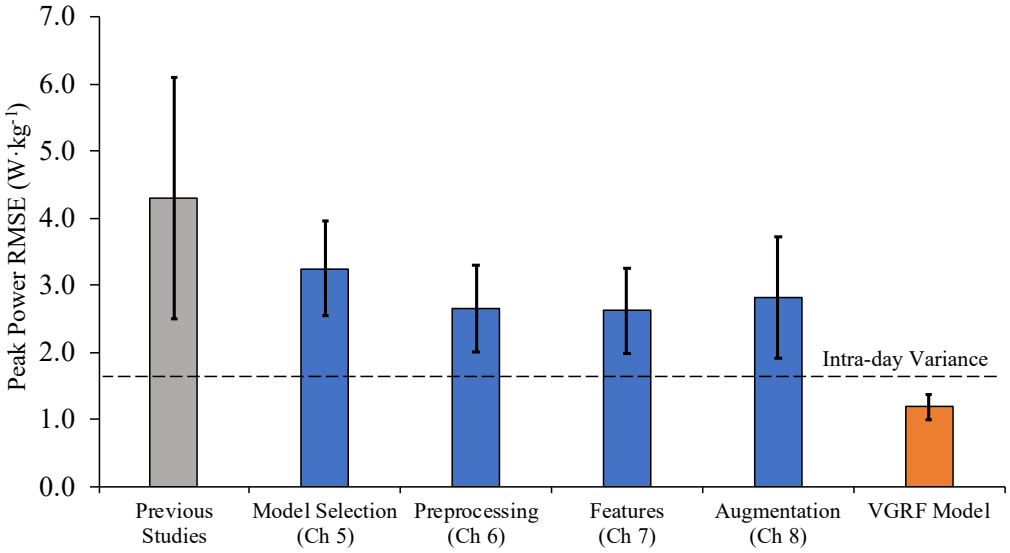


Figure 8-10. Progression in the GPR model's prediction errors through this thesis compared to previous studies (Ache-Dias et al., 2016; Amonette et al., 2012; Canavan & Vescovi, 2004; Lara et al., 2006; Quagliarella et al., 2010; Tessier et al., 2013). MCCV estimate shown for GPR with error bars based on the SD in validation error between folds. The intra-day variance is averaged from four reliability studies (Cormack, Newton, McGuigan, et al., 2008; Hori et al., 2007; McLellan et al., 2011; Taylor et al., 2012).

CHAPTER 9. GENERAL DISCUSSION

9.1 Introduction

The aim of this thesis was to develop a machine learning approach for predicting peak power in the countermovement jump based on accelerometer signals from a single body-worn sensor. It was the first biomechanical application of an FPCA-based model using body-worn accelerometer data. It was hoped that this machine learning approach might provide a low-cost, convenient alternative to the force platform gold standard. Overall, the accelerometer model achieved a predictive error of 2.5 $W \cdot kg^{-1}$, significantly improving upon previous methods for predicting peak power in vertical jumping. This final chapter provides formal answers to the individual research questions posed in Chapter 1. It takes a critical look at the methods employed, highlighting the value of this thesis, and reflects on the difficulties of using body-worn sensors to estimate performance metrics. Areas for further research are proposed.

9.2 Addressing the research questions

The research questions from Chapter 1 are answered in turn below, beginning with the two questions concerning accuracy before turning to the questions relating to various aspects of the model function that influenced its predictive error.

9.2.1 Model accuracy

I. How well does an FPCA-type model perform when predicting peak power and jump height in the CMJ, with and without arm swing, based on gold-standard VGRF data?

The models based on VGRF functional components performed well, explaining 98.3% of the variance in peak power and 92.0% of the jump height variance. The latter model was based on the work-done definition of jump height, but when the flight-time definition was used, the explained variance rose to 98.9%. In a further test of the FPCA technique, a logistic model correctly classified 88.8% of the jumps into those with and

without arm swing. This classification accuracy compared favourably with other activity recognition studies (Cust et al., 2019). In the case of jump height, r^2 was higher than previously reported in an FPCA study (79%), which was based on the flight-time definition (Richter et al., 2014b). The results showed that it was possible to achieve high r^2 without having to use ACP. Until this point, previous research had suggested only ACP models were capable of explaining a similarly high proportion of the jump height (Moudy et al., 2018; Richter et al., 2014a, 2014b). In conclusion, these results based on VGRF data indicated that FPCA would be a suitable feature extraction method for an accelerometer model predicting peak power. The merits of FPCA-type models and their errors will be discussed further in Section 9.3.2.

II. How well does an FPCA-type model perform when predicting peak power in the CMJ, with and without arm swing, based on accelerometer data from sensors at different anatomical positions?

The accelerometer model was developed in stages over Chapters 5–7 to ultimately yield a predictive error of 2.5 $W \cdot kg^{-1}$, equivalent to 5.8% at the mean of 45.0 $W \cdot kg^{-1}$. This was the error in peak power in the CMJ without arm swing (CMJ_{NA}), based on accelerometer data from the lower back (LB) sensor. It was larger than the 1.75 $W \cdot kg^{-1}$ *a priori* target based on the intra-day variance averaged from several studies (Cormack, Newton, McGuigan, et al., 2008; Hori et al., 2007; McLellan et al., 2011; Taylor et al., 2012). Although the model fell short of the accuracy needed for its intended application, it outperformed previous sensor-based methods for predicting peak power in vertical jumping. Those errors were two to eight times larger (RMSE = 4.9–21.6 $W \cdot kg^{-1}$; Table 2-2). The model compares favourably with the peak power prediction equations, based on body mass and jump height. The errors from those equations could differ considerably with different cohorts (2.0–27.6%; Section 2.3.2). The model's error was much smaller than the representative benchmark error from those studies, weighted by sample size, of 4.6 $W \cdot kg^{-1}$.

Chapter 5 showed that across a range of different algorithms, the errors for the CMJ with arm swing (CMJ_A) were consistently higher than their corresponding models for the CMJ_{NA}. The choice of jump type and sensor attachment position explained 92.3%

of the variance in the models' RMSE when averaged across the algorithms under investigation. This finding demonstrated the importance of data collection procedures and the choice of the sensor attachment site. The same analysis revealed that the errors were consistently larger for models based on the upper back and shank-attached sensors compared to the LB-sensor model. The decision was then made to focus the investigation on the LB-CMJ_{NA} data set given the consistent differential between the jump type and sensor-specific models. The differential between the LB- and UB-sensor models was not particularly large (1.0 $W \cdot kg^{-1}$), but it diminished the potential of re-purposing the UB sensors that are commonplace in professional sports teams. There would be obvious advantages of using the sensors the athletes already wear, as it would make field-based testing immediately convenient, as no setup would be needed. If the models could be made more accurate through future research, this small differential could be overcome. If so, it would considerably strengthen the case for a sensor-based system for predicting peak power in professional sport.

Taking a broader view, the final FPCA-based model holds up well to other models predicting various VGRF-related metrics. For comparison, the final model's predictions had a correlation of $r = 0.94$ with the true peak power values. From the literature (Table 2-4), studies reported correlations in jumping ($r = 0.73–0.94$, for peak VGRF and jump height) and in walking and running ($r = 0.50–0.86$, for peak VGRF and peak loading rate). This comparison excludes correlations of quantities averaged over multiple instances and subject-specific models. In terms of RMSE, only one model had lower percentage errors than the 5.8% error achieved in this thesis (RMSE = 3.8–5.0% for the peak VGRF in walking; Guo et al., 2017). However, caution should be exercised when making these comparisons as some outcome variables will be harder to predict than others, given the experimental conditions. With those caveats noted, it is still reasonable to conclude that the models developed in this thesis had an accuracy comparable to some of the best models and neural networks developed to predict VGRF-related measures, even with the strict unbiased cross validation procedures applied.

9.2.2 Model function specification

The model function encapsulated the full modelling procedure, including data augmentation, data preprocessing, feature extraction (FPCA), feature selection, model fitting and the inner cross validation loop. Its performance, through the optimisation procedure, provided the answers to the following research questions.

III. Which machine learning algorithms are well-suited to FPCA-based accelerometer data in regression models?

There were three models shortlisted in Chapter 5: regularised linear regression (LR), support vector machines (SVM) and Gaussian process regression (GPR). When data preprocessing was considered in Chapter 6, the SVM models were ruled out because they were unreliable, sometimes producing extremely large errors, while also being the most computationally intensive. Finally, the LR models were also ruled out as they did not benefit from feature extraction as much as the GPR models did (Chapter 7). Gaussian process regression was therefore the algorithm best suited to a model based on FPCA-type features when applied to the accelerometer data. This was also the first time a GPR model with its Bayesian approach was considered for a biomechanical application. Its success in this context should encourage other researchers to include it as a candidate model in their research.

IV. How should the accelerometer data be preprocessed to minimise the model's predictive error?

Preprocessing, as defined in this thesis, encompassed extracting the raw data from the appropriate time window, smoothing it with functional methods and warping the time domain with curve registration. These operations modified the accelerometer data without changing its form. The optimal time window included the jump, flight and landing, which marked a departure from the convention of jump VGRF models that takes data from jump initiation to take-off. The models made more accurate predictions due to the information available about the landing, mainly concerning its timing rather than the pattern of inertial accelerations following impact (Section 6.3.8). The optimal time window was not well-defined from the parameter distributions that emerged from

the optimisation procedure. Still, a reasonable inference was t_{pre} = 2000 ms and t_{post} = 1200 ms, with respect to take-off. The point of take-off (and landing) was identified using the accelerometer signal alone so the system could be self-sufficient, an essential requirement. It was accomplished to a higher degree of accuracy (± 12 ms) than in previous studies (Section 6.4.2). The hybrid approach combined the expertise in devising the algorithm with machine-led optimisation to tune its associated parameters (Section 6.2.1). It may serve as an exemplar for hybrid approaches in the future (Section 9.5).

The second preprocessing step was functional smoothing that took the place of conventional signal filtering techniques. It was optimal to smooth the acceleration data more heavily $(\lambda = 10^{4.8})$ than indicated by the standard fitting procedure based on generalised cross validation $(\lambda = 10^{2.0})$. As a regularisation parameter, the roughness penalty reduced the complexity of the acceleration curves, moderating the influence of higher-order FPCs (Section 6.4.7). Other studies have assumed a heavy degree of smoothing, implemented through signal filters with low cut-off frequencies, ≤ 3 Hz (Kautz et al., 2017; McGrath et al., 2019; Mlakar & Luštrek, 2017; Zago et al., 2019), but this is the first time functional smoothing has been used in this way.

Curve registration was an optional third preprocessing step that followed functional smoothing with the purpose of improving the curves' alignment. Registration had brought modest improvements to the VGRF model, but it could not do so for the accelerometer model. The VGRF model in Chapter 4 showed that it was essential to decompose the variance into amplitude and temporal components identified in the registration procedure. Having information about the time domain made the models more accurate, as was shown previously in a similar way with comparable ACP scores (Richter et al., 2014a, 2014b). Registration had not been similarly successful with the accelerometer models because the signals were so variable, making it difficult to identify a suitable set of landmarks that were unambiguous and present in all curves. Only the peaks in the pseudo power curves met this requirement, but the first peak could occur on either side of take-off, undermining the whole procedure.

V. What characteristics of the accelerometer signal are more important to the model in predicting peak power?

Taking the accelerometer signal in its resultant form yielded models with the lowest predictive errors. The resultant acceleration lacked directional information, but it was not confounded by sensor orientation changes, which in the CMJ were relatively minor for a lower-back mounted sensor. The resultant had a small advantage over the three-dimensional representation, which yielded models with a slightly higher error level (2.7 $W \cdot kg^{-1}$ vs. 3.0 $W \cdot kg^{-1}$; Table 7-3). The models based on the time-derivative and integrated curves achieved a similar predictive error (2.9 $W \cdot kg^{-1}$; Table 7-3). These results showed that the original acceleration was preferable to other data representations. The same conclusion was drawn concerning the varimax-rotated FPCs, which offered no advantage to the model but introduced unwanted multicollinearity.

Plotting the FPCs provided valuable insight into the nature of the component features. FPC3 exemplified this from the VGRF model (Figure 4-3C) that revealed the importance of a final surge in force production. It provided good experimental evidence supporting the principle of late force production, first proposed by Hochmuth and Marhold in 1977. The VGRF model also revealed how a combination of FPCs could produce unimodal and bimodal curves, providing a new perspective on the debate over the possible factors behind VGRF curve modality (Cockcroft et al., 2019; R. Kennedy & Drake, 2018; Lake & McMahon, 2018). This recent research is a return to the idea of the VGRF curve being a diagnostic tool (Luhtanen & Komi, 1978; Miller & East, 1976; Oddsson, 1987).

The accelerometer FPC1, which accounted for the largest proportion of the resultant curve variance, had the highest correlation with peak power. It described the temporal variance, reflecting variations in flight time, which was one of the two predictors in the peak power prediction equations. The FPCA-type models had achieved higher levels of accuracy than those equations by relying on additional components. FPC2, which had the second highest correlation with peak power, mainly described the variance in the impact acceleration spike's magnitude. Other FPCs were harder to

interpret from inspection as they became increasingly nuanced, even when revised with varimax rotations. However, the models utilised many components to make the most accurate predictions (12–29 FPCs).

VI. How does sample size and composition affect the model's predictive error, and can data augmentation bring worthwhile improvements in accuracy?

The model's predictive error declined with increasing sample size in an approximately logarithmic fashion with a slowing rate of improvement. At its full size, the training/validation data set with 275 jumps from 60 participants was not sufficient to achieve the target error of 1.75 $W \cdot kg^{-1}$, as discussed above. Therefore, it was necessary to estimate the likely sample size required based on an extrapolation of the model's learning curve (Figure 8-3). The best estimates suggested around 740 participants might be required, although the figure could exceed 1000 (Table 8-2). Nevertheless, it is clear that in practical terms, the final model developed in this thesis would not be capable of meeting the minimum level of accuracy. However, the analysis did show that the models had performed close to their full potential, giving reassurance that the investigation as a whole had sufficient data to draw valid conclusions.

The analysis also had broader implications for machine learning, as the resampling method revealed several facets of the model. The model learned more from different individuals than it did from additional trials from the same person, beneficial though they were (Figure 8-4). This evidence suggested the model benefitted from heterogeneity (a greater mix of participants), but the specificity principle still applied in the sense that errors were lower when there were more observations in a certain peak power range. In other words, the participants for training and validation should ideally have similar athletic abilities to those in the target population. If such volunteers are not available in sufficient numbers, data augmentation cannot fully compensate as the improvements in accuracy are relatively modest.

9.3 Methodological considerations

It is worth reviewing the methods employed, considering their strengths and limitations, and reflecting on their appropriateness to the research undertaken. This section will address the choice of sensor, FPCA for feature extraction, the novel optimisation procedure and the framework of nested cross validation, including the importance of the model function structure.

9.3.1 Sensor

In this thesis, the aim was to capture the characteristic patterns of movement from body-worn sensors in a new approach to estimating peak power output in vertical jumping. By identifying patterns instead of computing kinetic variables, the problems associated with the Newtonian methods could be avoided. In that sense, the orientation of a sensor was less of a concern, and so using a more sophisticated IMU coupled with an appropriate fusion algorithm was not required.

Accelerometers have been used more extensively than gyroscopes for activity recognition and predicting VGRF-related measures (Ancillao et al., 2018; Cust et al., 2019). Considering the nature of the CMJ, an in-place, explosive movement upward, it appeared that inertial accelerations would be the key factor in power production. Angular velocity would play a minor role, perhaps only in the rotation of the trunk in the sagittal plane through the propulsion phase. At the same time, the Delsys Trigno sensors offered the advantages of automatic and ongoing synchronisation of all sensors with the force platforms, solving a key technical challenge. As the subsequent analysis showed, the sensor's changing orientation was only a minor concern. The resultant signal yielded more accurate models than those based on the triaxial data, but only marginally so. When the FPCs based on the triaxial signal were used, the analysis showed that ten components related to approximately the vertical direction (X-axis) while two concerned the (approximate) anterior direction (Z-axis). Hence, the relevant patterns of inertial acceleration were predominantly along one axis, making the resultant a reasonable choice. This finding may not apply to other activities because the countermovement jump without arm swing is a well-controlled movement with limited degrees of freedom. The participants moved mainly with the same pattern,

greatly reducing the confounding effects of the sensor's changing orientation because all signals were affected in a similar way. This analysis was based on the LB-sensor data, but the Z-axis accelerations recorded by the UB sensor may play a greater role in an equivalent model. The UB sensor would experience greater tangential accelerations from the trunk's changing inclination during the jump, assuming the trunk behaves broadly as a rigid segment with the fulcrum at the hip joint. The UB-resultant acceleration may not be such a good representation of the movement as it was for the LB sensor, which may explain why UB-attached sensors were less accurate.

9.3.2 FPCA for feature extraction

FPCA was chosen as the feature extraction method because it had proven itself capable of identifying the characteristic features of time series data in many biomechanical applications (Section 2.5.2). The relevant literature also included studies specifically on the CMJ in which regression models had achieved high levels of accuracy in predicting jump height (Moudy et al., 2018; Richter et al., 2014a). As a data reduction technique, it could characterise a VGRF or acceleration curve with a small number of components. It made interpretation more intuitive, helping to reveal how the model worked rather than treating it as a black box. Various machine learning models could be trialled with moderate demands for data compared to neural networks. It was also computationally tractable to run a more extensive optimisation procedure with full nested cross validation. In short, FPCA had a significant bearing on the choice of subsequent methods.

FPCA depended on the curves having a high degree of alignment in the time domain. Aligning the curves at take-off was the most effective way of achieving this. This approach was particularly well-suited to the CMJ as the propulsion phase's timing is generally consistent. This approach may work well with other jumps and other explosive movements due to feed-forward neuromuscular control, reducing variability (Bobbert & van Ingen Schenau, 1988; van Soest et al., 1994). However, more generally, the versatility of FPCA is limited by its need for curve alignment, or in other words, its time invariance. Curve registration can partially realign curves based on common landmarks, as it did for the VGRF curves, but there was no satisfactory

solution for the acceleration curves. Registration was also computationally intensive, limiting its practical use even if suitable landmarks can be identified.

FPCA, and more generally PCA, is limited by being a linear decomposition. The projection onto a lower-dimensional subspace is defined by orthogonal hyperplanes, which further assumes the component scores have a Gaussian distribution (Bishop, 2006). These properties of PCA make it easy to fit and analyse the data, offering insights into the internal workings of the model, as was demonstrated in Chapters 6 and 7. But the principal components are unsophisticated features, lacking depth, and as such, they may not relate as well to the true characteristics of the data. Hence, the very low correlations with peak power output for all but the first two or three components (Figures 7–2, 7–3, 7–4). It also follows that the available features necessarily limit the complexity of ML models. Therefore, further improvements in model accuracy may come from nonlinear representations (Section 9.6).

9.3.3 Novel optimisation procedure

It was essential to find optimal values for the model function's parameters as its predictive performance was strongly dependent on them. The optimisation procedure incorporated three best-practice elements. It identified new observations with a random search, a more efficient technique for a multidimensional space than a grid search. From observations, the procedure constructed a Gaussian process model to represent the behaviour of the objective function (model function) using the same specification as that used by Matlab's Bayesian optimiser. The surrogate model's optimum was identified using the Particle Swarm algorithm, an efficient global optimiser.

It had become necessary to devise this approach because the Bayesian optimiser's results were not consistent when repeated with the same data set. The principal reason was that the objective function was noisy due to the sampling variance from the model function's MCCV procedure. It is a known issue with Bayesian optimisers, which was addressed here by conducting more extensive searches. Such an approach was only made possible by freeing the optimisation from the overhead of having to evaluate the computationally demanding acquisition function, whose cost grew exponentially with

each observation (Appendix E.5). The novel optimisation procedure instead relied on a random search that was progressively constrained to promising regions of the parameter space, guided by interim surrogate model predictions. As a result, the optimisation procedure could perform many times more iterations to pinpoint the global minimum more consistently with relatively moderate cost. One of its key benefits was the parameter distribution plots whose peak value identified the apparent optimal value (Cawley & Talbot, 2010; Krstajic et al., 2014; Varma & Simon, 2006), which was preferable to the bagged value when evaluated using the 1000-iteration MCCV procedure. In conclusion, the novel optimisation procedure produced more consistent outcomes at a moderate additional cost. It also revealed a greater insight into the model itself.

9.3.4 Nested cross validation

In this thesis, it was considered essential to develop a rigorous framework for selecting and appraising the models so the results would be representative of the model's performance on unseen data. The previous peak power prediction equations had not met this standard, even though some form of cross validation had been employed. Consequently, their reported errors were negatively biased (over optimistic), which only became apparent when those equations were tested independently on other groups. Nested cross validation was adopted for this investigation because it produces unbiased error estimates by separating model selection entirely from model appraisal (Section 2.6.2). It provided the structure that allowed repeated independent validations of the model with different subsamples of the data. Moreover, NCV appraised the whole modelling process and not just the model itself.

This thesis turned those ideas into a practical approach that was more extensive than in previous research. As defined in Chapter 1, the model function included all aspects of the modelling procedure, from data preprocessing and augmentation to feature extraction and selection, finally running the model itself in an inner cross validation loop. Whilst it is entirely possible and even likely that many others have created similar NCV implementations, NCV is relatively scarce in the literature despite its long understood advantages (Stone, 1974). The studies that have used it typically only

wished to demonstrate its capabilities with one or two parameters (e.g. Cawley & Talbot, 2010; Krstajic et al., 2014; Varma & Simon, 2006). In contrast, the model function in this thesis was driven by around 40 parameters in total, although only a few were manipulated at any one time. The model function was designed and implemented in Matlab to be driven by a global optimiser, such as the Matlab Bayesian optimiser, or by the novel optimisation procedure. It provides a framework for other researchers wishing to implement NCV.

Inevitably, the costs of running the model function were higher than the model itself due to the need to repeat the same preprocessing procedures that might otherwise have been carried out only once upfront. Nevertheless, the increase in computational cost was typically two to three times if augmentation is included. The higher cost became apparent when the model function was called thousands of times. Nevertheless, it is argued that these costs were justified in order to maintain the hermetic separation between model selection and appraisal so that the results had integrity. In this way, an effort was made to avoid making (unwittingly) over-optimistic claims of the accelerometer models' capabilities.

These results were supported by the holdout test that traditionally has been seen as the true independent test but is seldom undertaken in biomechanics research. Notwithstanding the merits of NCV, the holdout test perhaps still maintains its status because it so clearly and unambiguously is independent. It does not depend on trust that NCV is correctly implemented without inadvertently sharing data in some way. In this thesis, the holdout errors were quite similar to the NCV ones, indicating no such programming error. However, the holdout errors were less reliable because the test data set was small. It was clearly an independent test, but not necessarily a representative one. That is why the nested cross validation results should be preferable. Without the influence of selection bias, the errors may have been misleadingly smaller.

9.4 Challenge of sensor-based prediction

Aside from modelling considerations, it is worth reflecting on the challenges inherent in replacing a gold standard method with a less accurate alternative, especially one based on wearable sensors.

9.4.1 Errors in practice

In an athlete-testing regime, there will be a natural variation in test performances. The measurement error introduces another layer of variability on top, adding to the total measurement error, called the typical error in statistics (W. G. Hopkins, 2000). The implication of using an alternative method to the gold standard is that the typical error would be larger, reducing the test's sensitivity. In vertical jump testing, using a sensor-based method may be more convenient, but it would be harder to detect real changes in an athlete's peak power output, such as in response to a training programme or due to fatigue. What this means in practice can be understood by working through an example.

From Chapter 3, the within-subject variation for the CMJ_{NA} after taking account of known quantified factors was estimated at 4.0% (CV_{True}), based on the VGRF data. The variation arising from measurement error was 5.8% for the accelerometer model ($RMSE_{Model}$), so the observed coefficient of variation is:

$$CV_{Obs} = \sqrt{CV_{True}^2 + RMSE_{Model}^2} = \sqrt{4.0^2 + 5.8^2} = 7.0\% \approx 3.2\ W \cdot kg^{-1} \qquad (9.1)$$

Table 9-1 presents a series of examples of an apparent improvement in CMJ_{NA} peak power for an athlete based on two jumps, one before and one after a supposed training programme intended to increase power. In the pre-intervention test, their peak power is estimated at 45 $W \cdot kg^{-1}$ (also the mean value from the data collection). For the post-intervention jump, several hypothetical scenarios are considered where the estimated peak power ranges from 46 $W \cdot kg^{-1}$ to 53 $W \cdot kg^{-1}$. The significance of a change is whether it exceeds the smallest worthwhile change (SWC) that was estimated at 1.1 $W \cdot kg^{-1}$ (Chapter 3). Hopkins (2004) provided the formulae for calculating the probability of there being a substantial change (> SWC) and the rules-based inference.

Table 9-1. Illustration of the statistical inference that may be made on the measured change in peak power output between two CMJ tests based on either the best GPR accelerometer model or a force platform.

Peak Power	Sensor-based Model		Force Platform	
Change ($W \cdot kg^{-1}$)	Likelihood of Substantial Increase †	Inference ‡	Likelihood of Substantial Increase †	Inference ‡
45 → 46	49%	?	48%	?
45 → 47	58%	?	64%	?
45 → 48	66%	?	77%	↑↔
45 → 49	74%	?	87%	↑↔
45 → 50	80%	↑↔	94%	↑↑
45 → 51	86%	↑↔	97%	↑↑
45 → 52	90%	↑↑	99%	↑↑
45 → 53	94%	↑↑	100%	↑↑

† A substantial increase is defined as a greater than a 90% chance that the true change in peak power exceeded the SWC (Hopkins, 2004).
‡ Inference of a real change: ? Unclear; ↑↔ Possible increase; ↑↑ Very likely increase.
SWC = 1.5 $W \cdot kg^{-1}$; Typical Error = 3.3 $W \cdot kg^{-1}$ (Sensor-based Model); 2.1 $W \cdot kg^{-1}$ (Force Platform)
Typical error includes not just measurement error but many sources of performance variability.

Any increase in peak power recorded by the force platform may naively be taken at face value since it is the gold standard. Yet, in reality, the apparent improvement may be an illusion due to natural variations in performance from jump to jump. The analysis shows that no worthwhile change is discernible when using the accelerometer model until it exceeds 4.7 $W \cdot kg^{-1}$, but with a force platform, the difference may be larger than the SWC when the true increase rises above 2.2 $W \cdot kg^{-1}$. In contrast, a practitioner cannot be sure of the change (90% probability) until the true differences exceed 7.0 $W \cdot kg^{-1}$ and 4.4 $W \cdot kg^{-1}$, respectively. Thus, the sensor-based system is much less sensitive than the force platform to the actual changes, based on a one-off comparison. However, regular tests would be conducted in practice, which could mitigate their lack of sensitivity to some extent. These examples demonstrate how difficult it would be for a practitioner to have confidence in the sensor-based estimates of peak power based on a model with this level of accuracy. As Equation (9.1) indicates, reducing the model's error brings nearer the possibility of using sensor-based methods. However, there are two fundamental problems with sensor-based measurement that may limit further reductions in error that should be considered: the complications of movement variability and whether the sensor data contains the necessary information.

9.4.2 Movement variability

The degrees of freedom in the jumping movement allows many different ways to produce the same peak power output. Movement variability in jumping has been observed in terms of variation in the sequencing and timing of joint reversals in the shoulder, hips, knees and ankles (J. Jensen et al., 1989). Participants unconsciously prefer to employ a sub-optimal countermovement depth to retain flexibility in the system to respond to changing internal and external conditions (Bobbert et al., 2008). When recovering from heavy fatigue, athletes could return to nominal peak power levels by altering their movement patterns despite other measures indicating fatigue was still present (Gathercole et al., 2015).

Movement variability poses a problem for a sensor-based approach to testing because the sensor experiences greater variance in its movements than is present in the performance outcome. The locomotor system exploits the degrees of freedom available to achieve the high degree of control necessary for when it generates maximum force in the propulsion phase (Latash, 2012; van Soest et al., 1994). Consequently, for a sensor attached to a body part, variability may be even greater if the local dynamic behaviour is geared towards controlling the movement to fulfil a targeted objective (Bartlett et al., 2007; Davids et al., 2003; van Emmerik et al., 2016). Hence, sensor-based methods may be more susceptible to movement variability than force platforms depending on the anatomical location and its role in motor control. This point goes beyond traditional concerns of skin-movement artefacts in the data where the sensor moves relative to the skeleton underneath due to soft tissue movement and resonant oscillations, as discussed previously (Section 5.4.6). Movement variability was enough to discourage researchers in the past from using the VGRF profile as a diagnostic tool in jumping (Dowling & Vamos, 1993).

This situation still arose in a simplified jumping movement constrained by the participants placing their hands on their hips, despite jumping with arms being a natural choice. Efforts were made to counter these difficulties by identifying the best anatomical location for sensor attachment and the best jump type. Predictions of peak power were more accurate in the CMJ_{NA} than in CMJ_{A}, as expected. It was notable that heavily smoothing the accelerometer signal was advantageous as it may have

removed extraneous movements related to movement control rather than those directly related to the performance outcome. The model also relied on characteristics describing the flight time and landing, which are not connected to the variations in how peak power was achieved. However, the dynamics of landing would have incorporated variability, perhaps more so than in jumping. Despite such considerations, movement variability compounds the difficulty of translating those inertial accelerations into whole-body performance metrics. Further research is needed to investigate this issue further, as greater understanding may help inform machine learning methods in biomechanics.

9.4.3 Information in the signal

There is a more fundamental question as to whether the sensor data contains the information needed to determine the performance outcome. If it does, it would be theoretically possible to achieve the precision needed for predicting performance metrics, provided the technical means can be found. But if not, even the perfect model with its ideal feature extraction algorithm may not be able to achieve the precision required when determining peak external power or other such performance measures. This thesis and other studies (Table 2-4) have demonstrated with the models that there are complex relationships between inertial accelerations and various discrete VGRF measures. However, it is far from clear whether it will be possible to achieve the desired accuracy level for performance metrics in jumping or other activities based on body-worn sensor data.

It was noted from the outset that a sensor attached to a body segment could not follow the same trajectory as the whole body's CM. Thus, peak power estimates obtained with Newtonian methods could not be the same as the whole body's peak power output acting through the centre of mass, notwithstanding concerns over changing orientation and movement artefacts. The pattern-based approach in this thesis was more successful because characteristics in the recorded inertial acceleration bore relationships to peak power in various forms, which the GPR model combined appropriately. It picked out the indicators of higher-powered jumping, the most prominent of which were approximations of flight time and impact peak acceleration, as well as many others

that were harder for a human investigator to discern. Such characteristics appear to work well for classification problems as they act like signatures, distinguishing one class of activity from another. However, in the case of regression, the problem is significantly harder because it is not enough that those characteristics are present; they must also provide some indication of scale so that the model can compute the magnitude of the performance. FPCA worked well in this respect because its FPC scores provided a scale for the feature, translating to the inertial acceleration's magnitude. In summary, performance prediction from sensor data is a demanding problem, but ultimately accurate prediction may not be possible if the information is not present.

9.5 Practical applications of the research

Several key findings can be drawn from this thesis that may be of benefit to sports scientists and data scientists. For the purpose of athlete monitoring, there may be merit in classifying the accelerometer model's predictions into different bands rather than using the imprecise peak power values. This approach could be based on similar principles used above in Table 9-1 to classify the jump test result into no change, a significant increase or a significant decrease. It would be less sensitive than the force platform, but it would give the coach the information they would need with the desired convenience. It would still be a regression model, not a classification model because the performance classes would be individual-specific. All that would be required would be to add some logic to classify the regression model's estimate and present it to the user.

The results point to using the lower back as the preferred sensor attachment location rather than the upper back due to its proximity to the body's CM. This is an argument against taking advantage of the standard housing on the upper back commonly used for GPS tracking sensors. Caution should be exercised when working with measures obtained from IMUs housed in this way on the upper back. Although this finding is based on peak power output in the vertical jump, it may well have wider applicability because power generation itself is integral to explosive sporting movements. When

such considerations are made, it should be noted that data collection procedures, well designed and executed, have a much greater bearing on the model's output. Machine learning models depend on the sensor data containing the relevant information.

The resultant accelerometer signal can capture the movement's essential characteristics, despite the data not having directional information. Such information may contain spurious artefacts, so their absence helps to simplify the model to prevent overfitting. This recommendation is based on a controlled jumping movement with limited degrees of freedom in which individuals follow similar, regularised patterns of movement. In fact, many sporting movements could be described in this way where there is a commonly agreed template for the execution of the skill. This description could apply to running, kicking, throwing, rowing, cycling, swimming, and various types of jumps in volleyball and track and field (e.g. high jump, long jump, hurdling, etc.). In other situations where the movement is more irregular, IMUs may be more suitable to provide angular velocities or directional information (if fusion algorithms are sufficiently responsive).

FPCA can be a useful feature extraction method in situations that naturally place the time-series curves in close alignment with one another. Curve registration is quite limited in its effectiveness, especially with highly variable sensor data. Instead, aligning the curves at take-off in the vertical jump ensured close alignment in the period of interest when power output reached its peak. This approach would apply to other explosive movements where feedforward, self-organising mechanisms closely control the movement (Beek et al., 1995; Kelso, 1995; van Ingen Schenau, 1989; van Soest et al., 1994). It may also apply to cyclical movements, depending on their regularity.

It is recommended that the model function design framework should be adopted for model selection and optimisation as part of the nested cross validation framework. In this thesis, the AM function incorporated all aspects of data preparation, partitioning, modelling and (inner) cross validation. The whole modelling procedure should be assessed and optimised, not just the model itself. The novel optimisation procedure provided the means for doing so, achieving more consistent outcomes than the Bayesian optimiser. However, the parameters for optimisation should be limited to

prevent the search space from becoming too extensive. The parameters can be identified using a GLM from an initial survey. The GPR model, with its Bayesian approach, is also worth considering a candidate model, as it outperformed LR and SVM models in this study, two popular algorithms.

9.6 Future directions

FPCA-type models may be applied to other activities such as walking or running, having proved their worth with the CMJ, which served as an exemplar movement. There is considerable interest in peak forces and loading rates in human locomotion, such as in running, where the literature has focused on performance-related characteristics or indicators of injury risk (e.g. Cavanagh, 1987; Ferber, 2006; Nilsson & Thorstensson, 1989; van der Worp et al., 2016). The CMJ is a well-controlled movement with limited degrees of freedom, but other dynamic movements may be much more challenging. In situations where curve alignment cannot be so easily achieved, acceleration time series could be represented as a linear combination of wavelets. Wavelet scattering is a time-invariant method that can adapt to features positioned at different points on the curve (Andén & Mallat, 2014; Mallat, 2012). Its flexibility may be more suitable for highly variable accelerometer signals.

This thesis has thoroughly investigated FPCA as a feature extraction method, but future research may need to turn to other methods to achieve higher levels of accuracy. Popular as it is in biomechanics and effective as it has been shown to be for machine learning, FPCA lacks the sophistication of state-of-the-art techniques. PCA methods project data onto fixed linear manifolds (hyperplanes), but real data often reflect the influence of low-dimensional nonlinear manifolds (Bishop, 2006). Without such flexibility, the encoded features may not fully reflect the latent characteristics of the system. Independent Component Analysis (ICA) provides a step-up in sophistication, allowing components to be blended together without the requirement for orthogonality or a Gaussian distribution, but ICA has been superseded by autoencoders thanks to significant advances in recent years (Creswell et al., 2018; Goodfellow et al., 2016).

Autoencoders (AE) are feedforward neural networks that perform a nonlinear dimensional reduction by passing data through a bottleneck in the (undercomplete) network topology (Hinton, 1990). The decoder following the bottleneck reconstructs the data from the encoded features. The encoded representation has been shown to yield a smaller reconstruction error than PCA when using the same number of dimensions (Hinton et al., 2006). Moreover, by channelling data through a bottleneck, the network has to learn a lower-dimensional encoding that captures the essential characteristics of the data. Thus, AEs could be a powerful alternative method for feature extraction for the sensor data. Research should determine the best encoder design for data in this context, considering the depth of the network and various techniques to guard against overfitting, such as denoising, regularisation and the use of sparsely activated nodes (Goodfellow et al., 2016).

Autoencoders may be considered a form of unsupervised learning, as with PCA, in the sense that the encoding is made irrespective of the outcome variable. As a result, the encoded features may not be an ideal representation for the problem in question. Alternatively, DNNs are supervised as they are trained to minimise the outcome prediction error, and so the network learns to encode features tailored to the specific problem. A recurrent network design such as an LSTM would appear to be a sensible choice for a time series input. LSTMs have been applied before to inertial sensors for activity recognition where they outperformed non-recurrent DNNs (Ordóñez & Roggen, 2016). No prior feature extraction method would be required since the network would perform its own feature encoding internally. However, early investigations with LSTMs for this thesis did not yield promising results with raw or smoothed accelerometer data as inputs. A better architecture design is required that will need to be determined by a suitable optimisation procedure such as Bayesian optimisation or genetic algorithms, which has been employed successfully for DNNs (Bouktif et al., 2018; Domhan et al., 2015; Olson et al., 2016). Although LSTMs may be a reasonable first choice, non-recurrent networks such as convolutional networks (CNNs) are still worth considering. The time-dependence element may be less important once the signals have been aligned in wearable sensor applications, as was

the case in this thesis. Other studies for activity recognition applied CNNs to a fixed time window of accelerometer data (Kautz et al., 2017).

Although AEs and DNNs have greater sophistication, they typically have many hundreds or thousands of hyperparameters that require tuning, pushing up the demand for higher data volumes. Nevertheless, more advanced data augmentation methods have emerged based on generative models that can produce realistic synthetic data (Creswell et al., 2018). Variational Autoencoders (VAEs) draw from the encoded features' distribution to construct new data (Kingma & Welling, 2014). In principle, these methods should work well because proximity in latent space corresponds to similarity in the data space, unlike with FPC space due to the descending importance of each dimension (component). Thus, research should investigate how much more effectively augmented sensor data can be produced using VAEs. Essentially, it would mean replacing one feature representation for a more sophisticated one but applying the same principles described in Chapter 8.

Synthetic data can also be produced using Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), which can produce remarkably realistic data. GAN research initially focused on image processing and natural language processing, but recurrent GANs for time series data have now been developed (Wen et al., 2021; Yoon et al., 2019). Conditional GANs (CGANs) can produce synthetic data for a given class, which is suitable for classification problems (e.g. Smith & Smith, 2020), but generating data associated with a continuous variable has been a more fundamental problem. However, new research has proposed regression CGANs, the most promising technique being feature contrasting (Aggarwal et al., 2020; Olmschenk et al., 2019; Rezagholiradeh & Haidar, 2018). Thus, it is now possible to produce synthetic data for a desired value of the continuous outcome variable. With few applications thus far, there is a considerable opportunity to develop this technique for inertial sensor data. Hence, it should be possible with further research to produce accelerometer signals for a given peak power value. These synthetic data should be of a much higher quality than was possible in Chapter 8 and thereby help improve the model's predictive accuracy. However, a note of caution is needed as training GANs can be problematic due to instability and non-convergence. Nevertheless, a consensus is emerging on the

best approaches, with many examples of how researchers have stabilised their GANs (Wiatrak et al., 2020).

Future research could investigate using AEs to predict the VGRF profile rather than discrete performance measures as the shape and magnitude of the force-time curve is as important as the precise value at any one point. There has been considerable work in this area, but the problem is not yet solved (Section 2.4.22.4.2). Autoencoders may offer a more elegant solution by transforming accelerometer data to VGRF data. Typically, AEs are trained with a loss function measuring the dissimilarity between input and output, where the latent feature encoding is of interest. However, if the loss function were redefined as the RMSE with the true VGRF, or some other suitable measure of dissimilarity, the AE would learn a suitable transformation function. Hence, using the acceleration curve as input, the AE could produce the corresponding VGRF curve output. The same principles underpin transformation functions in image processing (e.g. colourisation) (e.g. Isola et al., 2018) and neural language translation (e.g. Cheng, 2019). An exploratory investigation before writing this thesis demonstrated that a feedforward network could transform accelerometer FPCs scores into VGRF FPC scores. Despite being based on a simple linear encoding, the reconstructed VGRF curves had a high degree of fidelity ($r = 0.96$; mean RMSE = 0.05 BW). Future research could take a more sophisticated approach by developing AEs to perform this task. This idea could be taken further by using conditional AEs that factor in additional inputs such as the relevant participant characteristics – for example, body mass or limb lengths may be relevant factors. This transform could also be reversed so that an AE could use VGRF data to generate synthetic accelerometer data. Once the transform is learned, the AE could be used for data augmentation by exploiting the extensive archive of VGRF data accumulated from previous research.

In summary, there are several promising areas for further research to investigate how state-of-the-art techniques can be applied to inertial sensor data in biomechanics applications. FPCA worked well for the CMJ, a well-controlled movement with limited degrees of freedom, but as a linear decomposition, it had its limitations. More complex movements are likely to require more sophisticated latent representations if

their associated performance measures can be determined with a sufficient level of accuracy. Hence, there is a considerable opportunity for biomechanical applications of machine learning. Although research has been dedicated mainly to classification problems, the field is still relatively immature compared to other disciplines such as image processing and natural language processing that were first to embrace these new techniques. Regression models are now rising to prominence in these fields, with enormous potential for applied biomechanics practice. There are new challenges in learning how best to harness these techniques to biomechanical problems, not least in developing optimal network architecture designs. For instance, GANs can suffer from instability and fail to learn. Their first use could be in generating highly realistic synthetic data to augment datasets and help improve model accuracy. But in a fast-moving field, there may be many more innovative applications.

9.7 Conclusion

This thesis investigated a new method of predicting peak power in the CMJ using body-worn accelerometer data as a low-cost, field-based alternative to the force platform gold standard. Feature extraction was based on FPCA, a technique that has had notable success in a diverse range of biomechanical applications but hitherto has not been employed with machine learning models other than in simple linear regression. The models were more accurate than previously published methods based on equations using body mass and jump height, and Newtonian methods applied to sensor data. The final model also compared favourably to other biomechanical applications predicting VGRF-related discrete measures. Ultimately, the model could not achieve the accuracy required to make them a practical proposition for athlete testing. Nonetheless, this thesis has developed new approaches to optimisation and developed a more comprehensive framework for nested cross validation, the first to do so in biomechanics. With new nonlinear encoding techniques, there is a considerable opportunity to advance biomechanics practice further based on the similar approaches. The many techniques advanced in this thesis can be taken up by other researchers in their quest to develop sensor-based models for performance-measure prediction. It is

hoped this work will help the field continue to progress towards more accurate sensor-based solutions.

APPENDIX A. DATA COLLECTION METHODS

This appendix presents further details on the data collection.

A.1 Participants

Table A-1. Participants assigned to the training/validation data set.

ID §	Sex	Age (yrs)	Height (m)	Body Mass (kg)	Primary Sport (Secondary)	Primary Sport Level	Protocol †	No. Jumps ‡
1	M	24	1.75	62.4	Running	Club	A	8
2	M	21	1.75	75.9	Rugby (Cycling)	Club	A	8
4	M	22	1.84	76.5	Surfing	Recreational	A	8
5	M	20	1.80	90.9	Rugby	Club	A	8
6	M	20	1.80	59.8	Running	Recreational	A	8
8	M	22	1.91	87.8	Football (Basketball)	Club	A	8
9	M	27	1.85	99.5	Basketball	Club	A	8
10	M	20	1.79	83.1	Football	Club	A	8
12	M	21	1.82	84.4	Rowing	National	A	8
13	M	21	1.87	66.6	Tennis	Club	A	8
14	M	19	1.98	98.8	Volleyball	Club	B	16
15	F	28	1.66	64.3	Running	Recreational	A	8
16	M	35	1.88	94.8	Ice Hockey	Club	A	8
19	M	24	1.77	72.4	Football	Recreational	A	8
20	F	23	1.69	62.9	Swimming	Club	A	8
21	M	27	1.76	98.6	Gymnastics (Rugby)	National	A+B*	16
26	M	18	1.89	79.8	Football	Club	A	8
27	M	19	1.81	71.7	Football	Club	A	8
29	M	23	1.59	66.4	Kickboxing	Club	A	8
30	F	33	1.77	69.9	Netball	Club	A	8
31	M	18	1.81	66.1	Sailing	National	A	8
32	M	19	1.88	94.4	Badminton	Club	A	8
34	M	19	1.75	68.7	American Football	Club	A	8
35	M	22	1.76	90.1	Triathlon	National	A	8
37	F	22	1.67	55.2	Gym	Recreational	A	8
38	F	19	1.62	49.3	Running	Club	A	8
39	M	21	1.83	74.6	Rock Climbing	Recreational	B	16
44	M	19	1.77	63.8	Football	Recreational	A	8
47	F	20	1.65	61.9	Kickboxing	National	A	8
49	M	19	1.80	93.2	Rugby (Taekwondo)	Club	A	8
50	F	18	1.40	51.4	Tennis	Club	A	8
53	F	19	1.57	52.0	Gym	Recreational	A	8
56	M	19	1.81	66.9	Cricket (Running)	Club	A+B*	16
57	F	20	1.74	58.1	Netball (Running)	Club	A	8
58	F	20	1.67	68.4	Volleyball	Club	A	8
60	F	25	1.75	68.9	Weightlifting	Club	A	8
61	F	21	1.65	63.2	Gymnastics (Sprinting)	National	A	8
63	M	20	1.72	62.4	Tennis (Long/Triple Jump)	Recreational	A+B*	16
64	F	20	1.57	54.6	Kickboxing	National	A	8
66	M	19	1.94	72.0	Football	Club	A	8
68	F	23	1.77	100.0	Rowing	Club	B	16
69	M	19	1.68	73.0	Gym	Recreational	A	8
72	M	20	1.90	91.1	Rowing	Club	A	8
73	F	25	1.65	70.8	Netball	Club	A	8
74	M	21	1.65	73.5	Rowing	Recreational	A	8

ID §	Sex	Age (yrs)	Height (m)	Body Mass (kg)	Primary Sport (Secondary)	Primary Sport Level	Protocol †	No. Jumps ‡
75	M	23	1.66	56.2	Swimming	Club	A	8
76	M	21	1.85	97.1	Rowing	Club	A	8
77	F	23	1.68	66.6	Hockey	Recreational	A	8
79	M	28	1.70	56.6	Volleyball	National	A	8
80	M	19	1.73	81.6	Cricket (Jiu-Jitsu)	Club	A	8
83	M	18	1.79	69.0	Hockey	Club	A	8
84	M	22	1.78	77.2	Rugby	Club	A	8
86	M	20	1.84	89.4	Football (Swimming)	Club	B	16
87	F	21	1.62	60.1	Netball	Club	B	16
90	F	18	1.61	74.6	Volleyball	Club	A	8
91	F	21	1.65	69.8	Volleyball	Club	A	8
94	F	26	1.58	62.4	Gymnastics	Recreational	A	8
96	M	21	1.75	68.0	Volleyball	Club	A+B*	16

† Protocol A = 8 CMJ (4-4 split with/without arms) + 8 broad jumps; Protocol B = 16 CMJ (8-8 split) ; Protocol A+B* (for returning participants) where B* = 8 CMJ (4-4 split). ‡ vertical jumps (CMJ_{NA}+CMJ_{A}). § Participants were assigned a random ID between 1 and 99, which accounts for gaps in the sequence.

Table A-2. Participants assigned to the holdout testing data set.

ID §	Sex	Age (yrs)	Height (m)	Body Mass (kg)	Primary Sport (Secondary)	Primary Sport Level	Protocol †	No. Jumps ‡
11	F	23	1.68	68.0	Netball (Cross Fit)	Club	B	16
22	F	26	1.68	66.1	Hockey	National	B	16
28	M	25	1.74	85.3	Rugby	Club	B	16
40	M	19	1.67	68.5	Judo (Swimming)	National	B	16
43	M	18	1.87	77.4	Swimming	National	B	16
82	M	23	1.73	76.4	American Football	Club	B	16
88	M	18	1.74	68.5	Football	Club	B	16
95	F	25	1.69	87.6	Football	Recreational	B	16
97	M	20	1.75	75.2	Basketball	Recreational	B	16

† Protocol A = 8 CMJ (4-4 split with/without arms) + 8 broad jumps; Protocol B = 16 CMJ (8-8 split) ; Protocol A+B* (for returning participants) where B* = 8 CMJ (4-4 split). ‡ vertical jumps (CMJ_{NA}+CMJ_{A}). § Participants were assigned a random ID between 1 and 99, which accounts for gaps in the sequence.

A.2 Force platforms' calibration

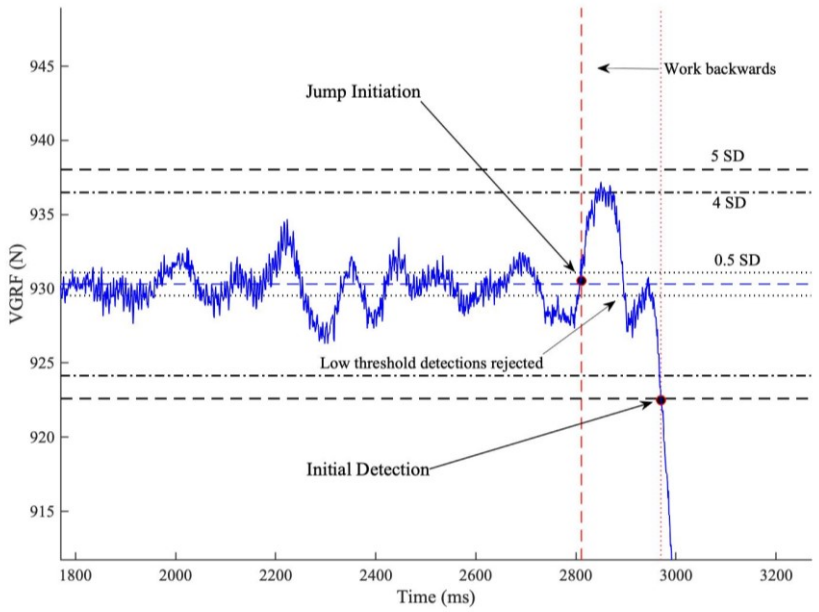
Two portable force platforms were calibrated using certified weights to produce criterion VGRFs of 0 N, 100 N, 200 N and 300 N. The VGRF measurements were averaged over a suitable steady-state period of 1–3 s. With both force platforms combined, the calibration measurements extended up to 600 N. A calibration equation for each platform on each day was formulated using a linear best fit based on the set of four criterion and practical VGRF measurements. The correlation coefficients were near perfect: $r = 1.0000$, to four decimal places in every case. The random errors for each force platform were 0.28 ± 0.32 N and 0.32 ± 0.32 N, respectively, when averaged across all days of testing.

A.3 Jump detection algorithm

The following algorithm identifies the jump initiation point in the VGRF time series. It is adapted from the method developed by Owen et al. (2014) to handle the variety of VGRF patterns seen in 500+ jumps. Owen et al. (2014) defined the initial detection of movement as the first point where the magnitude of the net VGRF exceeded: $\left|F_{\text{Net}}(t)\right| > 5 \times \text{SD}$. The SD was taken over the first 1 s of measurement. However, the movement would have begun slightly earlier than the detection point, so Owen et al. (2014) assumed it was 30 ms earlier. Instead of making an assumption, an algorithm was devised for this thesis to find a more appropriate point using the VGRF curve.

The algorithm worked backwards from the initial detection point to find the first point when $\left|F_{\text{Net}}(t)\right| < 0.5 \times \text{SD}$ (Figure A-1A). (By working backwards, the inequality is reversed.) This smaller threshold was possible because movement had already been detected. A false detection would be when $F_{\text{Net}}(t)$ passes through zero, as illustrated in Figure A-1B, defined by $\left|F_{\text{Net}}(t)\right| > 4 \times \text{SD}$ over the preceding 100 ms. If it did, the algorithm repeated the same procedure from the new earlier point; if not, the point was accepted. The algorithm produced backwards time offsets from the initial detection point of 70 ± 53 ms across all jumps, indicating this more adaptable approach was appropriate.

(A)



(B)

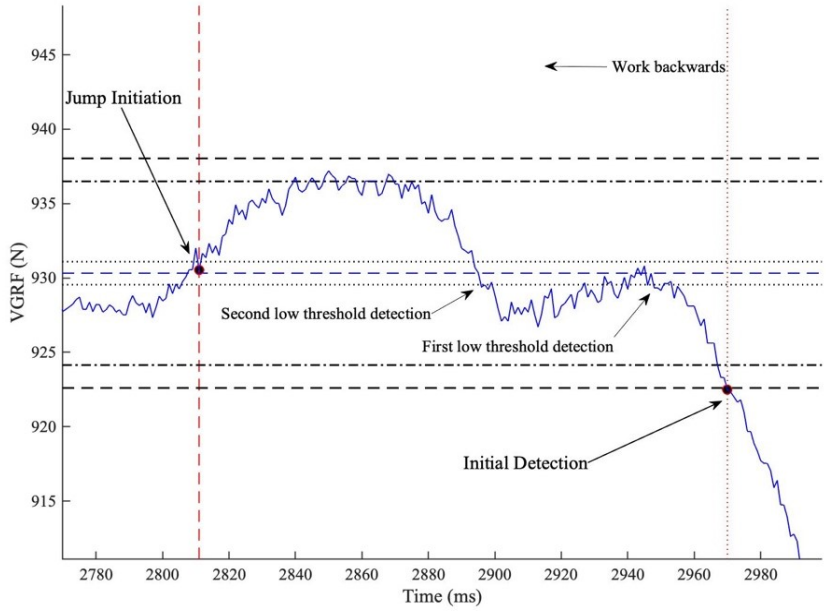


Figure A-1. Jump Initiation Algorithm – an example. Note the granular scale on the y-axes. (A) The initial detection of movement is made when the VGRF deviates from the estimated BW by > 5 SD. The BW is the mean VGRF over the first second. The SD is the standard deviation of VGRF over this period. The first detection of movement is not the best estimate for the timing of jump initiation. The algorithm works backwards to find where the VGRF has deviated by less than 0.5 SD. (B) This candidate for jump initiation at t = 2953 ms must be verified by working backwards further from this point to determine whether the VGRF remains within a 4 SD threshold for a minimum of 100 ms. In this example, it does not and passes below the 0.5 SD threshold another time (t = 2899 ms) before a valid jump initiation point is accepted at t = 2811 ms.

APPENDIX B. SENSOR SELECTION

This appendix describes the selection of a suitable make and model of the sensor from the laboratory's inventory. This involved considering the requirements for such a sensor followed by an investigation into their validity and reliability. For the purposes of comparison, a simple scoring system was devised to rank the sensors based on how well they matched up with the requirements.

B.1 Requirements

The following technical and functional requirements were considered in view of the intended application to detect vertical jumping movements. It was envisaged that the sensors would be attached directly to the skin of the participant at anatomical locations commonly used for wearable sensors in sport: the upper and lower back and the lower leg (Section 3.2.4). The following requirements were drawn from the literature, as well as from the manufacturers' documentation.

- **Onboard sensors** – triaxial accelerometers were required that measure linear acceleration accurately along orthogonal axes. Some sensors may also include gyroscopes, magnetometers which come under the more general category of inertial measurement units, but those capabilities were outside scope.
- **Mass** – the sensors should be as light as possible, preferably much less than 20 g, so they can be attached to the skin or clothing without adding an appreciable mass to the body or interfering with movement. A light sensor also has a higher resonant frequency, especially when compression is applied using strapping or bandages (Saha & Lakes, 1977).
- **Sensitivity Range** – the accelerometers should have a linear response across the range of accelerations experienced during the movements under investigation. Preliminary investigations with the countermovement jumps, with and without arm swing, suggests that ±10g would be sufficient if the sensors are not attached to the wrists.

- **Sampling Frequency** – a high sampling frequency was required to record the jumping movement in sufficient detail. Based on conventional approaches for time normalisation, the sample should contain 100 points as a minimum, although more is desirable. Since countermovement jumps typically require ~1 s from the first movement to take-off (Robertson & Fleming, 1987), 100 Hz may be considered the lowest acceptable sampling frequency.

- **Calibration** – all acceleration measurements must be calibrated since accelerometers produce an output voltage proportional to the inertial acceleration. The sensor control software may offer automated calibration as a feature, but otherwise, manual procedures will be required.

- **Acceleration validity** – the measurement of acceleration at each time interval should be accurate, as expressed by the validity correlation. The validity correlation should exceed 0.995, which is classed as "excellent" (W. G. Hopkins, 2015). The random error should be less than the smallest meaningful difference.

- **Temporal validity** – the accuracy of the times logged should be very high to ensure that measurements remain synchronised with other sensors. It is conventional in time measurement to state accuracy in terms of the accrued error ("drift") as the variance in clock cycles every million clock cycles, expressing the error as parts per million (PPM) (Lombardi, 2008). For instance, a clock that drifts one second every month has an error of 0.4 PPM. Over the data collection session (< 1 hour), the drift should be less than the smallest meaningful time difference. The time resolution from the sampling frequency also imposes a lower limit on this meaningful difference. With a sampling frequency of 100 Hz, the time resolution would be 10 ms. Hence, for the drift to be less than 10 ms over one hour, the systematic error should be less than 2.8 PPM.

- **Temporal reliability** – if the sensors' measurement of time drifted from the actual time at a uniform rate (which may be specific to each sensor), then the drift could be corrected systematically. The piezoelectric crystal, which regulates the internal clock, oscillates with a resonant frequency that can vary to a small extent due to imperfections, temperature changes, mechanical

stresses and the electrical circuit design (Vig, 2000). These variances should amount to no more than 10 ms or no more than one time interval.

- **Synchronised data logging** – the study requires multiple sensors to record accelerations concurrently, remaining synchronised with minimal time drift for the data collection session. Sensors that individually record data internally depend on their internal clocks' accuracy to remain in sync with one another. Other sensor types transmit their data to a central data logging station ensure their data remains synchronised.

- **Simultaneous initiation** – the sensors start logging data at the same time as each other. Some systems allow the user to configure and start multiple sensors at the same time.

- **Operating range** – for sensors that transmit data to a central data-logging unit, the operating range becomes an important practical consideration. Jump testing takes place within the confines of the laboratory, where an operating range of 10 m is sufficient.

- **Data series partitioning** – the start time and end times from each trial should be identifiable in the acceleration-time series, clearly and unambiguously, to facilitate the analysis. This requirement is met by default by systems that support individual recordings for each trial. For those that do not, a bespoke method would be required to identify each jump. Hence, data recorded in separate trials would have significant practical advantages.

- **Systems integration** – sensor data will need to be cross-referenced with data from force platforms, requiring some level of integration between the systems. Integration may also facilitate synchronisation.

- **Data access** – when dealing with accelerometers, occasionally, data turns out to have issues associated with it. If this only becomes apparent afterwards when the data is uploaded, the trial data may have to be discarded. However, for sensors that transmit their data continuously, issues may become apparent immediately so they can be resolved there and then during the data collection.

B.2 Sensor types

The selection process considered three MEMS-type sensors (Micro-Electronic Mechanical Systems), representative of the range of types of sensors available.

- **Actigraph GT9X Link monitors** (Actigraph Corp., Pensacola, FL, USA) were developed primarily for activity tracking and clinical and movement skills assessments. They include an accelerometer, gyroscope and magnetometer with a maximum sampling frequency of 100 Hz. The vendor's ActiLife software configures and starts the sensors simultaneously while they are connected to the computer. The sensors calibrate themselves automatically and log data independently of each other. Hence, they depend on their internal clocks to stay in sync with one another. The sensors must be reconnected to the computer to upload their data once data collection is complete.

- **Wildbyte activity sensors** (Wildbyte Technologies Ltd., Swansea, UK) were developed for tracking animal behaviour and migratory patterns. They include an accelerometer and magnetometer as well as GPS, temperature and pressure sensors. The accelerometer's sampling frequency can be set up to 800 Hz with other options at 160 Hz, 80 Hz and 40 Hz. As with Actigraph, the sensors record their measurements internally and can only be uploaded later, one sensor at a time. The operator must start each sensor separately using a microswitch. There is no automated calibration feature.

- **Delsys Trigno sensors** (Delsys Inc., Natick, MA, USA) were developed primarily to measure the electrical activity of skeletal muscle (electromyography or EMG), but they also include triaxial accelerometers. The sensors transmit their readings to a Delsys base station continuously, which collates the data, ensuring it is synchronised. The vendor's EMGWorks software can process the digital signals, but other systems can access analogue systems, allowing accelerometer data to be combined and synchronised with other data sources. Each trial can be recorded separately by starting and stopping data capture.

B.3 Assessment

A points-based system was used to assess the three sensor types above. A priority (high, medium, and low) was assigned to each requirement. Points were assigned based on how well each sensor type met the requirement: 3 points for high compliance, 2 points for medium and 1 point for low.

B.3.1 Preliminary Assessment

A preliminary assessment was made without considering temporal validity, temporal reliability and acceleration validity (Table B-1). The Delsys sensors meet the majority of the requirements and achieve the highest score. They offer practical advantages over the other two types in terms of bespoke sampling frequencies, data synchronisation and integration, real-time data presentation, and the partitioning of data collections into separate trials. However, they also have some limitations. With a range of ± 9g, the accelerometer is unlikely to be able to track rapid movements when the accelerometers are placed distally on the limbs where linear acceleration would be highest. Their operating range is limited to 20 m from the portable base station, which would impose restrictions on how and where the sensors can be used. For this selection exercise, the cost was not considered a factor because all three sensor types were part of the laboratory's inventory.

Table B-1. Preliminary assessment of the three sensor types. The importance of each requirement is indicated by its assigned priority. If the sensor type meets the requirement, indicated by an underline, the score for that requirement is added to the sensor type's total. The requirements for temporal validity, temporal reliability and acceleration validity were not included.

	Requirement	Priority (Score)	Delsys	Actigraph	Wildbyte
Onboard Sensors	Accelerometer	High (3)	Accelerometer EMG	Accelerometer Gyroscope Magnetometer	Accelerometer Magnetometer
Mass	< 20 g	Medium (2)	13.2 g	16.1 g	6.5 g
Sensitivity Range	≥ ± 10g_0 [b]	High (3)	± 9g_0	± 8g_0, ± 16g_0 [c]	± 16g_0
Sampling Frequency	≥ 100 Hz	High (3)	250 Hz,(user-defined) [a]	30-100 Hz	5-800 Hz
Calibration	Automatic	Low (1)	Manual	Automatic [d]	Manual
Synchronised Data Logging	Synchronised	High (3)	Synchronised	Independent	Independent
Simultaneous Initiation	Synchronised	Medium (2)	Synchronised	Synchronised [e]	Independent
Data File Partitioning	Individual Trials	Medium (2)	Individual Trials	All Trials [f]	All Trials [f]
Operating Range	10 m	Low (1)	20 m	Unlimited	Unlimited
Systems Integration	Integrated	Medium (2)	Integrated [g]	None	None
Data Access	Real-time	Low (1)	Continuously	Upload	Upload
	PRELIMINARY SCORE		**19**	**15**	**12**

(a) For analogue signals– digital signals are fixed at 148.15 Hz; (b) Depends on context and anatomical location; (c) Secondary accelerometer; (d) Single value calibration: Z = 1g_0, X = 0, Y = 0; (e) Up to 6 sensors at one time; (f) Individual trials must be identified and extracted; (g) With Vicon Nexus 2.5

B.3.2 Assessment of Temporal Validity and Reliability

The sensors' temporal validity and reliability were assessed in an investigation that compared the timing accuracies of vibrations generated by forceful impacts. The sensors detected the vibrations, as did a force platform which served as the criterion method (Appendix C). The findings were as follows:

- The systematic timing errors (drift) were 1.2 PPM for Delsys (3.1 s per month), 1.5 PPM for Actigraph (3.9 s per month) and 11406 PPM (8.2 *hours* per month) for Wildbyte sensors. When corrected for drift, the random errors were 25 ms for Delsys, 21 ms for Actigraph and 124 ms for Wildbyte.
- The Delsys and Actigraph sensors had high reliability equivalent to a variance of less than one time interval every 10 seconds. In terms of absolute time, the Delsys sensors had a smaller variance of 2.5 ms compared to 3.2 ms for Actigraph. The Wildbyte sensors were much less reliable, with timing variances equivalent to 7 time intervals over 10 seconds, or 28.4 ms in absolute time.
- The Delsys sensors could run for over an hour, while the Actigraph sensors could run for around 50 minutes before temporal drift exceeded the tolerance level of 10 ms. The Wildbyte sensors' unreliability was such that they could only be run for a little over 5 minutes before the timing variance became too great.

The Wildbyte sensors fell far short of the requirements and were excluded from the second part of the investigation.

B.3.3 Assessment of Acceleration Validity

The acceleration validity investigation involved attaching the sensors to a spinning wheel that produced accelerations that could be measured accurately with a motion capture system (Appendix D). It established that the Delsys and Actigraph sensors both achieved excellent validity overall ($r = 0.998$). Nonlinearities were low for both sensor types, although the Actigraph errors were slightly larger at higher magnitudes. The

random errors of the Delsys sensors were lower than those for the Actigraph sensors (0.061g vs. 0.100g). The Delsys sensor had marginally the better validity, but either sensor type could be used for the study.

B.3.4 Conclusions

This selection exercise aimed to determine which of the three available sensor types would be best suited for recording participant movements in vertical jumps. The preliminary assessment above indicated that the Delsys sensors offered the best overall fit based on their technical specifications and functional capabilities. The Wildbyte sensors could not meet the temporal requirements. The investigation into acceleration validity found no clear difference between the Delsys and Actigraph sensors. In the final assessment (Table B-2), Delsys and Actigraph meet most of the requirements, but the Delsys sensors come out ahead overall. The stand out advantages of the Delsys sensors was the automatic synchronisation of accelerometer signals from all sensors and the ability to also synchronise them with the VGRF data from force platforms. Synchronisation also made it possible to capture separate recordings of accelerometer and VGRF data obviating the need to extract the jump measurements from long sensor recordings. The Delsys sensors were therefore chosen for the jumps study in this thesis.

Table B-2 Final assessment of the three sensor types that include temporal validity and reliability, and acceleration validity. The mean of the main result is quoted for each category along with the score in brackets. Scoring: 3 = meets or exceeds the requirement; 2 = close to the requirement; 1 = falls short of the requirement; 0 = clearly fails to meet the requirement.

		Requirement	Delsys	Actigraph	Wildbyte
		PRELIMINARY SCORE	19	15	12
Temporal Validity	Systematic Error	≤ 2.8 PPM	1.2 (3)	1.5 (3)	11406 (0)
	Random Error	≤ 10 ms	2.1 (3)	2.5 (3)	12.4 (2)
	Max Running Time	≥ 1 hour	1 hr 3 min (3)	50 min (2)	5 min (1)
Temporal Reliability	Inter-sensor Variance	≤ 1 time interval	0.8 (3)	1.3 (2)	7.1 (1)
Acceleration Validity	Validity Correlation	≥ 0.995	0.998 (3)	0.998 (3)	n/a (0)
	Dynamic Random Error	≤ 0.01g	0.061g (3)	0.100g (2)	n/a (0)
	FINAL SCORE		**37**	**30**	**16**

APPENDIX C. SENSOR TEMPORAL VALIDITY AND RELIABILITY

This appendix describes the investigation into the temporal validity and reliability of three types of sensors as part of the sensor selection exercise. Appendix B sets out the requirements and describes the three sensor types under investigation: Delsys, Actigraph and Wildbyte.

C.1 Methods

The approach taken was to compare the timings of impact vibrations that would be detectable by the sensors and force platforms almost simultaneously. Six sensors of each type, 18 in all, were placed on an MDF wooden board (340 mm × 270 mm × 17 mm, 0.98 kg) on top of a Kistler force platform embedded into the floor (model 9281EA, 600mm × 400 mm; Kistler Instruments UK Ltd., Hook, UK) – Figure C-1. The force platform (1000 Hz) was connected via a Kistler control unit (model 5233A) and an analogue-to-digital converter.

Figure C-1. The sensor layout on a wooden board, which was placed on top of the force platform. Top row: Actigraph sensors; middle row: Delsys sensors; bottom row: Wildbyte sensors.

Timing signals were generated by striking the board with one end of a steel bolt (mass 157 g). The vibration from the impact would be detected almost instantaneously by the sensors and the force platform underneath (<0.025 ms based on the speed of sound through hard wood). The vibration signals in the VGRF were the criterion measure of time, which was assumed to have a temporal error of 1 PPM, the typical standard for computer-grade piezoelectric crystals (Marouani & Dagenais, 2008). Timing signals were produced approximately every 10 seconds over 15 minutes, making 90 signals in all. The impacts were evident in the acceleration and force traces as sudden sharp spikes. The time of each impact was when the force/acceleration rose above the noise level. The timing resolution was therefore governed by the sampling frequency (i.e. force platform = 1000 Hz, Delsys = 250 Hz, Actigraph = 100 Hz, Wildbyte = 160 Hz).

C.1.1 Temporal validity

Linear regression models were developed for each sensor to predict the criterion time (from the force platform) based on the sensor's timing. Each model was based on a set of 97 criterion and sensor times. The systematic error, equivalent to the timing drift, was defined as the slope minus one, expressed as a percentage. With perfect agreement – no drift – the slope of the best fit straight line would be one. The random error was defined as the standard error in the estimate (SEE). The random error's confidence interval was based on a chi-squared distribution as the lower bound is zero. The intercept gave the predicted time when the sensor's time was zero. The intercept was the signal processing delay for the Delsys sensors, which had started simultaneously with the force platform. All confidence limits were set at 90%.

Using the above linear models to predict the correct time may not eliminate the timing drift entirely because there is uncertainty in the systematic error's precise value. If the systematic error estimate differs from the actual drift, there would be a growing divergence between the corrected time and the criterion time. When the cumulative drift exceeds the smallest meaningful time difference (10 ms), the sensor would need re-synchronisation. This point is effectively the maximum running time (MRT). It was estimated for each sensor by dividing the smallest meaningful difference by half the confidence interval of its systematic error. The representative MRT for the sensor type

was defined as the lower confidence limit (i.e., the shorter time) of the mean MRTs of all six sensors. The mean MRT's confidence limits were based on a chi-squared distribution since the MRT is bounded at zero, i.e. it must be positive.

The models' residuals, the differences between predicted times and the actual times, increase over time. This cumulative error inflates the random error, but it is not an artefact of the sensors or the experimental method. A second linear regression was performed to remove each sensor's cumulative error by using the time gaps between successive timing signals. Thus, all times were around the 10-second mark, the typical timing gap between strikes of the steel bolt, which more closely reflects the anticipated times required to record the movements under investigation.

C.1.2 Temporal reliability

The inter-sensor reliability was determined using the sensors' measurements of the timing gaps above. It was equivalent to a test-retest reliability analysis to find the intra-class correlation ICC(3,1), based on the 15 possible combinations of sensor pair comparisons for each timing gap. The random errors were halved because they included the uncertainty from two timing points, not one (W. G. Hopkins, 2000). The thresholds for ICC were set at: 0.20 low, 0.50 moderate, 0,75 high, 0.90 very high and 0.99 extremely high (W. G. Hopkins, 2015). It was not possible to test the reliability of individual sensors because the exact timing gaps between each strike with the steel bolt were not consistent.

C.2 Results

The temporal validity correlation was the highest for the Delsys sensors, followed by the Actigraph sensors and Wildbyte sensors (Table C-1). The systematic and random errors of the Delsys and Actigraph sensors were both low, whereas those of the Wildbyte sensors were five orders of magnitude higher. When the random error calculation was based on time gaps, which eliminates the inflation in random error, the random error was lower, substantially so for Wildbyte. Based on the uncertainty in the systematic error value, the elapsed time required for the timing drift to exceed the chosen tolerance of 10 ms was around one hour for Delsys, 50 minutes for Actigraph, but just over 5 minutes for Wildbyte. The Delsys signal processing delay was 102 ms, with a 90% confidence interval of ±1 ms.

Table C-1. Timing accuracy by sensor type showing the mean with a 90% confidence level.

	Delsys	**Actigraph**	**Wildbyte**
Validity Correlation	0.999999999947	0.999999999917	0.999999988639
Systematic Error (%) [a]	0.00012 ± 0.00017	0.00015 ± 0.00022	-1.14064 ± 0.00241
Systematic Error (PPM)	1.2	1.5	11406
Random Error (ms) [b]	2.9 ± 0.4	3.6 ± 0.4	40.4 ± 4.9
Random Error (ms) [c]	2.1 ± 0.3	2.5 ± 0.3	12.4 ± 1.5
Max Running Time [d]	1 hr 3 min	50 min 8 sec	5 min 35 sec

a. Systematic error represents the rate of drift over time
b. Random error based on absolute times.
c. Random error based on the gaps between timing signals.
d. Maximum running time is the expected time it would take for drift to exceed the smallest meaningful difference (2 ms). The lower confidence limit is quoted.
The results highlighted in bold are the best for that quantity.

The Delsys sensors had the highest inter-sensor reliability, followed by Actigraph and then Wildbyte (Table C-2). The timing variance between sensors of the same type can be expressed in seconds and time intervals, which considers the sampling frequency and better reflects its practical importance. Using the Delsys timing variance as the baseline, the ‘true error’ for Actigraph and Wildbyte showed the effect of individual sensor timing differences alone without variance arising from the method.

Table C-2. Timing reliability between sensors of the same type showing the mean with a 90% confidence interval. More decimal places than usual are quoted to differentiate between each sensor type.

	Delsys	Actigraph	Wildbyte
ICC (inter-sensor)			
(based on time)	0.99999 ± 0.00001	0.99997 ± 0.00001	0.99902 ± 0.00039
(based on time intervals)	0.997 ± 0.001	0.992 ± 0.003	0.764 ± 0.079
Error between sensors			
(time in ms)	3.2 ± 0.4	5.2 ± 0.6	28.4 ± 3.5
(time intervals)	0.8 ± 0.1	1.3 ± 0.2	7.1 ± 0.9
True error †			
(time in ms)	-	1.9 ± 0.6	25.1 ± 3.5
(time intervals)	-	0.5 ± 0.1	6.3 ± 0.8

† With Delsys as the baseline, thereby eliminating the error in identifying the times of the vibration signals.
The figures highlighted in bold were the best in each row.

C.3 Discussion

The aim of this investigation was to determine the temporal validity and reliability of the Delsys, Actigraph and Wildbyte sensors as part of the sensor selection exercise described in Appendix B. The results show that the Delsys and Actigraph sensors achieved high levels of accuracy and inter-sensor reliability. In contrast, the Wildbyte sensors failed to meet any of the requirements, specifically: a systematic error ≤ 2.8 PPM, a random error ≤ 10 ms, and an effective maximum running time exceeding one hour.

The Delsys and Actigraph sensors achieved a very low systematic error in time measurement equivalent to a drift of approximately one second per week. This error compares relatively well with a typical wristwatch that may typically lose one second per month (Lombardi, 2008). In contrast, the Wildbyte sensors had a systematic error that was five orders of magnitude larger. The considerable uncertainty in the precise value of Wildbyte sensors' systematic error resulted in a short running time of around 5 minutes before re-synchronisation would be required. Such short running times would make experiments with Wildbyte sensors impractical as data collections would

have to be frequently interrupted to re-synchronise the sensors. The Wildbyte sensors' substantial random errors contributed to their lower temporal validity.

The inter-sensor reliability test had a similar outcome with the Delsys and Actigraph sensors, achieving consistency, while the Wildbyte sensors showed much greater disparities. The Wildbyte sensors could differ from each other by as much as seven time intervals after only 10 seconds of running. A countermovement jump may be short (i.e. 1-2 s for the jump itself), but the sensor would have to be running since the start of the data collection, which could last 30 minutes or longer. The cumulative error by this point could be quite large (1,260 time intervals equivalent to over 7 s, based on the above drift). The uncertainty in the timing would make comparisons between accelerometers unreliable. The piezoelectric quartz crystals that govern the Wildbyte sensors' internal clock may be of a lower quality. They have a rating of 20 PPM, according to their design engineer (personal communication). Furthermore, the Wildbyte sensors log data to mini SD cards, which as a type of flash memory is known to have slower, more variable write times.

It may be thought that since the Delsys sensors are synchronised with each other, there should be no timing difference between them. The fact that a difference exists (up to 10 ms) suggests a degree of unreliability in the synchronisation method. The method depends on detecting the first appearance of a signal amidst the variable noise. Accounting for this error by making Delsys the baseline, it becomes apparent that the Actigraph sensors demonstrate a high degree of consistency with each other despite their independent internal clocks.

The intercept time from the linear regression for the Delsys sensors was equivalent to the delay resulting from the signal processing. The time of 102 ms is close to the Delsys base-station's processing delay of 96 ms, as reported by the manufacturer. A further delay of 6 ms may occur during the analogue-to-digital conversion performed by the Vicon Giganet box. The overall delay of 102 ms is effectively fixed (with a random error of 1 ms or less), allowing it to be used as a fixed timing offset when using Delsys accelerometers with Vicon in future studies.

APPENDIX D. SENSOR ACCELERATION VALIDITY

This appendix describes the investigation into the validity of the Delsys and Actigraph sensors' measurement of inertial acceleration. Appendix B describes the sensor selection exercise, of which this is part. An earlier investigation into the sensors' temporal validity and reliability found that the Wildbyte sensors could not meet the requirements, excluding them from this investigation into acceleration validity.

D.1 Method

D.1.1 Experimental Setup

Six sensors of each type (Delsys and Actigraph, 12 in all) were attached to a bicycle wheel (700C) with the frame turned upside down and tilted away from the vertical. The tyre and inner tube were removed so the sensors could be attached directly to the metal rim that was made from an aluminium alloy (Figure D-1). The Delsys and Actigraph sensors were attached alternately to the rim with a regular 10° separation using white insulating tape. The triaxial accelerometers were aligned with the wheel's radial, tangential and perpendicular axes. Three possible attachments were tried but only with the sensors laid flat across the rim could correct alignment be ensured.

The wheel was stationary for a static measurement for calibration when the sensors only detected gravity. The wheel was then spun by hand and left to spin for 50 s so that the sensors experienced inertial accelerations on top of the gravitational acceleration. The radial acceleration was principally due to the centripetal force from the wheel's rotation. Tangential accelerations were induced by attaching the sensors over a 120° arc to unbalance the wheel, causing it to slow down and speed up. A constant perpendicular acceleration was induced by tilting the wheel away from the vertical.

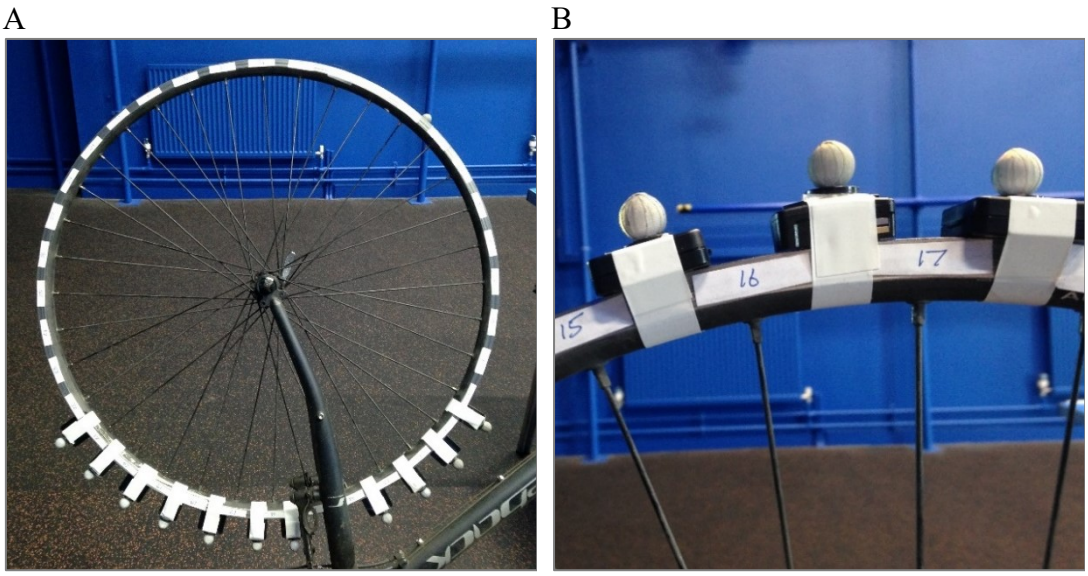


Figure D-1. Acceleration validity experimental setup. (A) The inverted bicycle wheel with sensors attached with reflective markers on top. (B) Alternating sensor attachment with insulating tape with Actigraph sensors on either side of the central sensor.

D.1.2 Motion capture

The movements of the sensors were recorded using a motion capture system (T20S, Vicon Motion Systems Ltd., Oxford UK) based on reflective markers that were attached to the top of the sensors. Twelve T20S cameras tracked the sensors' movements with a frame rate of 250 Hz, the same as the Delsys sensors. The marker trajectories were reconstructed using Vicon Nexus 2.5 software without using the automatic gap-filling functions available.

The Actigraph sensors were synchronised with Vicon by striking the rim with a rigid, plastic ruler three times. The vibrations were transmitted around the rim and via the spokes, reaching the sensors < 0.32 ms after impacts, based on the speed of sound through aluminium. A reflective marker attached to the end of the ruler allowed the Vicon cameras to track the ruler's movements. The average timing of the three impacts (measured by Vicon and by each of the sensors) were used to shift the time series into line with one another.

D.1.3 Criterion acceleration calculation

The radial acceleration experienced by a sensor is the combination of the centripetal acceleration and the component of gravitational acceleration in that direction:

$$a_R = \omega^2 r + g_R \tag{D-1}$$

where ω is the wheel's angular velocity and r is the radius of the arc followed by the accelerometer inside the sensor. g_R is the radial gravity component which is defined further below. The tangential acceleration combines the changes of angular velocity with the gravitational acceleration component:

$$a_T = \alpha r + g_T \tag{D-2}$$

where α is the wheel's angular acceleration and g_T is the tangential gravity component.

The sensor positions estimated by Vicon were converted from Cartesian to cylindrical polar coordinates (Figure D-2A). The x and z coordinates were swapped so that the angular coordinate θ represented the sensor's angular position, with 0° being vertical (Figure D-2B). The origin was found using the theorem that the perpendicular bisectors of two chords intersect at the circle's centre. The normal to the circle and hence the wheel's orientation is the cross product of the chords. The origin and normal vectors' calculations were based on three positions of the same sensor at different time intervals separated by approximately one third of a revolution and averaged over ten revolutions.

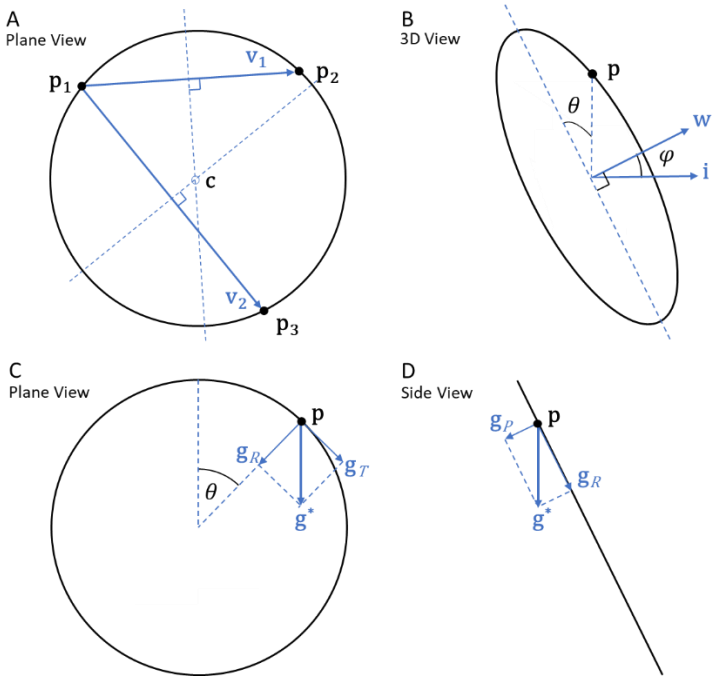


Figure D-2. Geometry of an inclined circle in 3D. (A) perpendicular bisectors of two chords meet at the centre, c. The centre is a linear combination of the chord vectors, v_1 and v_2. (B) the circle plane is rotated such that its normal vector, w, aligns with the x-axis. (C) the gravity vector has components radially and tangentially, where θ is the angular coordinate. (D) the gravity vector also has a perpendicular component.

The calculations depended on accurately determining the axel's position, the wheel's orientation, the radii of the sensors' circular paths and whether the sensors correctly lay on the wheel's plane. Their associated measurement errors were estimated from data taken from 100 revolutions (Table D-1). The errors were smaller than the positional accuracy of the Vicon system itself. The tilt angle was estimated as 23.2°. The sensors' circular path was calculated to have radii of 341.75 ± 0.65 mm for Delsys and 337.67 ± 0.75 mm for Actigraph.

Table D-1. Positional measurement errors represented by their standard deviations. The errors are less than the overall accuracy of the motion capture system, which is attributed to improving the accuracy of calculations based on circular geometry.

Positional Measurement	Standard Deviation
System position accuracy †	± 0.88 mm
Centre position variance ‡	± 0.39 mm, ± 0.69 mm, ± 0.86 mm
Sensor out-of-plane variance	± 0.61 mm
Wheel inclination variance *	± 0.120°
Radius variance	± 0.70mm (0.21%)

† Based on a calibration using the Vicon wand that found the mean distance error between the wand's 5 markers (10 marker-pair combinations). The wand was swept at random within the Vicon system's capture volume during a 40-second sampling period.
‡ Typical variance for a given sensor (the centre was calculated individually for each sensor)
* With respect to the wheel's normal axis, which is also equal to the variance with the vertical

The gravity components radially, tangentially and perpendicular are defined as follows:

$$g_R = g_2^* \sin\theta + g_3^* \cos\theta \tag{D-3}$$

$$g_T = g_2^* \cos\theta - g_3^* \sin\theta \tag{D-4}$$

$$g_P = g_1^* \tag{D-5}$$

where (g_1^*, g_2^*, g_3^*) is the gravity vector in the local coordinate system of the sensor (Figure D-2C & Figure D-2D).

The criterion acceleration for each sensor was found using the above expressions and Equations (D-1) and (D-2), based on the marker positional data. All calculations were made using custom-written scripts in Matlab R2018a. The acceleration due to gravity was taken to be 9.812 ms^{-2} (latitude 51.6°, altitude 10 m) based on the International Gravity Formula (Moritz, 2000).

The angular velocity and acceleration due to the rotation must be the same for all sensors at each instant in time due to the wheel's circular geometry (excluding gravitational acceleration). Therefore, rather than calculate the angular velocity and acceleration independently for each sensor, accuracy was improved by basing the calculations on the time-varying mean angular position of all 12 sensors. The mean

angular position was differentiated using the central difference formula to find the angular velocity, which was differentiated a second time to find the mean angular acceleration. The angular velocity and angular acceleration were smoothed using a Gaussian-weighted moving average with a window of 20 time intervals. The gravitational acceleration vector was different for each sensor as it depended on the angular position. An individual sensor's angular position was found relative to the collective mean angular position of all sensors by applying the individual sensor's mean angular offset.

The sensors' accelerometers were calibrated with the wheel held in a series of 18 stationary positions, each separated by approximately 20°. The accelerations recorded would be based on gravity alone. A calibration equation was determined using linear regression for each axis and each sensor. A further linear regression using data from the rotating wheel was used to determine the correlation between the calibrated acceleration measurements and the criterion accelerations. The validity correlation, which is the Pearson correlation, was evaluated based on the following scale: 0.70-0.85 poor, 0.85-0.95 good, 0.95-0.995 very good, > 0.995 excellent (W. G. Hopkins, 2015). The criterion acceleration was calculated at 250 Hz as it was based on marker positions, matching the sampling frequency of the Delsys sensors. Compared with the Actigraph sensors running at 100 Hz, the criterion acceleration was resampled at 100 Hz using cubic spline interpolation.

D.2 Results

The Delsys and Actigraph sensors' measurement of acceleration had an excellent level of validity (> 0.995) along the radial axis, but along the tangential axis, validity was lower ('very good') (Table D-2). The Actigraph sensors achieved the higher accuracy of the two in the tangential direction. The random error was consistently lower for the Delsys sensors in both axes compared to the Actigraph sensors.

Table D-2. Sensor validity from the dynamic trial.

		Delsys	Actigraph
Validity [a]	Radial	0.9984 [0.9983, 0.9986]	0.9980 [0.9977, 0.9983]
	Tangential	0.9754 [0.9731, 0.9776]	0.9806 [0.9776, 0.9832]
Random Error (g_0)	Radial	0.061 ± 0.002	0.100 ± 0.005
	Tangential	0.014 ± 0.000	0.022 ± 0.001

Figures highlighted in bold are the best for that quantity.

The sensors were subject to a range of inertial acceleration over the 50 s of the dynamic trial, which saw the highest acceleration, as expected, along the radial axis, up to ~ 9.7g (Table D-3). The tangential accelerations reversed direction as the unbalanced wheel's angular velocity varied. The perpendicular acceleration was virtually constant, which is why no validity figures can be presented for it.

Table D-3. Accelerations during the dynamic trial. Mean ± 90% CI shown, minima and maxima in square brackets.

	Delsys	Actigraph
Tangential Acceleration (g_0)	-0.016 ± 0.536 [-0.616, +0.523]	-0.014 ± 0.537 [-0.613, +0.541]
Perpendicular Acceleration (g_0)	0.393 ± 0.000 [-]	0.393 ± 0.000 [-]
Radial Acceleration (g_0)	7.792 ± 1.059 [5.845, 9.769]	7.697 ± 1.050 [5.765, 9.669]

Both sensor types demonstrated a very low level of nonlinearity (< 1%) with increasing accelerations, based on a log-log transformation of the radial acceleration prior to the linear regression (Table D-4). This transformation allows the random error to be expressed as a percentage, which is lower for the Delsys sensors.

Table D-4. Linear regression based on a log-log transformation for the radial acceleration only. The mean errors with the 90% CI are expressed as percentages.

	Delsys	Actigraph
Systematic Error	0.46 ± 0.13%	0.57 ± 0.33%
Random Error	1.66 ± 0.06%	2.68 ± 0.14%

Figures in bold are the best for that quantity.

D.3 Discussion

The aim of this investigation was to determine the acceleration validity of the Delsys and Actigraph sensors as part of the exercise to select the type of sensors to use for this thesis. The assessment of validity was considered in two ways: (1) the correlation between the measurements made by the sensors and the criterion acceleration based on the motion capture data – this is conventionally thought of as the primary measure of validity; and (2) the random error in those measurements defined as the standard error of the estimate from the linear best fit of the data. The Delsys and Actigraph sensors were rated excellent for their acceleration measurements along the radial axis, but along the tangential axis, Delsys was rated only as 'very good', while Actigraph maintained its excellent rating. The lower levels of validity in the tangential axis are likely to be due to misalignments about the radial axis, which will be discussed in more detail below.

It was easier to differentiate between the sensors based on the random errors, for which the Delsys sensors had consistently lower errors than Actigraph. The Delsys sensors' higher sampling rate may account for this based on simulations comparing best fits of time series sampled at different rates. The higher random errors in the accelerometers aligned with the radial axis are expected since those accelerations were much higher. Both types of sensor had very low nonlinearities (< 1%), indicating that their measurements of high and low accelerations would be equally valid. In this respect, Delsys was marginally better.

In the experimental setup, the sensors lay flat across the wheel's rim. It was evident that there was good alignment of the sensors' Z-axis with the radial axis thanks to the

rim. However, confidence was lower in the alignment of the sensors' X-axis with the tangential direction and of the Y-axis with the perpendicular. Although care was taken, a sensor could have turned away very slightly from pointing directly along the rim even though the insulting tape held them firmly in what should be the correct alignment. The sensors were re-attached flat to the rim after being turned 90° to swap the X and Y-axes around. The data from this arrangement was not reported for brevity, but the figures were similar and the conclusions the same. The X and Y axes were also tested when they were aligned with the wheel's radial axis. Therefore, conclusions on the validity of the sensors' acceleration measurements must be drawn from only the Z-axis being subject to high accelerations. It must be assumed that each onboard accelerometer has the same level of accuracy. The data from this and other trials do not provide evidence to the contrary.

When the validity figures, the random errors and the nonlinearity levels are considered as a whole, the Delsys sensors come out marginally ahead. In practice, there is very little to tell between the Delsys and Actigraph sensors. Both sensor types meet the acceleration validity requirements for this thesis. The choice of which sensor to use must be based on other considerations discussed in Appendix B.

APPENDIX E. ACCELEROMETER MODEL

This appendix presents the analyses supporting the methods developed in Chapter 5 for the accelerometer model. It includes the results determining the number folds and iterations for the K-fold design for the grid searches and the MCCV design. It also reveals the bias introduced when the data is partitioned by trial rather than by participant. Finally, it provides evidence of the Bayesian optimiser's inconsistency and its cost, which grows exponentially with the number of iterations.

E.1 K-Fold Cross-Validation Design

Cross-validation provides a representative estimate of the model's accuracy across multiple sub-samples of the data. The k-fold design can be tailored either for model selection – discriminating between models based on their loss – or for model appraisal – determining the best estimate of the model's predictive error. For model selection, it is advantageous to minimise the variance in the loss between folds to make it easier to discriminate between models. For model appraisal, the opposite is the case where the aim is to minimise bias. These two divergent aims reflect the bias-variance trade-off. The investigations in Chapters 5, 6, 7 and 8 first focused on model selection before switching to model appraisal to better estimate the selected model's predictive error. Chapter 2 previously reviewed the literature for the recommended approaches to cross-validation. This section now lays out the supporting evidence for the number of folds chosen and the number of iterations performed based on the data gathered for this thesis.

The LB-CMJ$_{NA}$ data set was used for this investigation, which evaluated the 11 models introduced in Section 5.2.3 (Table 5-1), providing a broad range of models for the assessment. The k-folds considered were {2, 5, 10, 60}, the latter being equivalent to the Leave-One-Out (LOO) cross-validation, all of which performed at the participant- rather than the trial-level. The cross validation design was based on the 20-iterations MCCV procedure with 2, 5, 10, 20, 50 and 100 folds.

The results showed that the validation loss (RMSE), averaged across all models, declined for more k-folds, a sign of reducing bias (Table E-1). At the same time, increasing the number of iterations led to an asymptotic approach to the true loss estimate for each k-fold design, the best approximations being the 100-iteration estimates. The same relationship was observed for all models (Table E-2), which is not universally the case (Burman, 1989; Kohavi, 1995; Luo, 2016; Y. Zhang & Yang, 2015).

Table E-1. Average validation loss, defined at the RMSE ($W \cdot kg^{-1}$), for different combination so k-folds and iterations, averaged across all 11 models.

	Iterations					
K-Folds	**2**	**5**	**10**	**20**	**50**	**100**
2	4.37	4.37	4.39	4.43	4.47	4.48
5	4.50	4.49	4.48	4.46	4.41	4.36
10	4.27	4.26	4.26	4.24	4.22	4.19
60	4.12	4.11	4.10	4.09	4.05	4.00

Table E-2. Validation loss, RMSE ($W \cdot kg^{-1}$), for different k-folds over 100 iterations for individual models.

	Models (100 Iterations)										
K-Folds	**LR**	**RDG**	**LSS**	**TREE**	**BST**	**SVM-L**	**SVM-G**	**GPR-SE**	**GPR-M52**	**NN-5**	**NN-10**
2	3.51	3.43	3.43	6.31	4.94	5.52	5.57	3.89	3.90	4.16	4.88
5	3.37	3.30	3.32	6.17	4.68	5.11	5.12	3.66	3.72	3.93	4.55
10	3.29	3.27	3.21	6.01	4.55	4.88	4.94	3.51	3.65	3.74	4.30
60	3.18	3.16	3.21	5.79	4.46	4.59	4.68	3.36	3.46	3.30	3.76

LR = Linear Regression; RDG = Ridge regression; LSS = Lasso regression; TREE = Single tree; BST = Tree ensemble using the Boost method; SVM-L = Support Vector Machine with a Linear kernel; SVM-G = Support Vector Machine with a Gaussian kernel; GPR-SE = Gaussian Process Regression with a Squared Exponential kernel; GPR-M52 = Gaussian Process Regression with a Matérn 5/2 kernel; NN-5 = Feedforward Neural Network with 5 hidden nodes; NN-10 = Feedforward Neural Network with 10 hidden nodes.

The standard deviation in the validation loss was determined for each iteration count (N) using the formula: $SE_N \approx SD_{100}/\sqrt{N}$, where SD_{100} represented the 100-iteration case considered to be a good approximation of the true variance. The SE indicated the uncertainty in the estimate of the loss across all models. Averaged across all models, the SE was higher for more k-folds, a sign of increasing variance, whilst it declined for larger numbers of iterations due to the inverse relationship with N (Table E-3). The same relationship can be observed in all models (Table E-4).

Table E-3. Standard error (W·kg^{-1}) for different combinations of k-folds and iterations, averaged across all 11 models.

	Iterations					
K-Folds	**2**	**5**	**10**	**20**	**50**	**100**
2	0.37	0.23	0.16	0.12	0.07	0.05
5	0.46	0.29	0.21	0.15	0.09	0.07
10	0.55	0.35	0.25	0.17	0.11	0.08
60	0.88	0.56	0.39	0.28	0.18	0.12

Table E-4. Standard error (W·kg^{-1}) for different k-folds over 100 iterations for individual models.

	Models (100 Iterations)										
K-Folds	**LR**	**RDG**	**LSS**	**TREE**	**BST**	**SVM-L**	**SVM-G**	**GPR-SE**	**GPR-M52**	**NN-5**	**NN-10**
2	0.34	0.34	0.32	0.74	0.49	0.60	0.69	0.53	0.38	0.46	0.81
5	0.46	0.45	0.45	0.76	0.61	0.86	0.91	0.63	0.52	0.63	0.92
10	0.62	0.55	0.56	0.93	0.74	0.98	0.99	0.72	0.67	0.89	0.95
60	1.03	1.06	1.07	1.40	1.27	1.54	1.56	1.06	1.06	1.28	1.41

LR = Linear Regression; RDG = Ridge regression; LSS = Lasso regression; TREE = Single tree; BST = Tree ensemble using the Boost method; SVM-L = Support Vector Machine with a Linear kernel; SVM-G = Support Vector Machine with a Gaussian kernel; GPR-SE = Gaussian Process Regression with a Squared Exponential kernel; GPR-M52 = Gaussian Process Regression with a Matérn 5/2 kernel; NN-5 = Feedforward Neural Network with 5 hidden nodes; NN-10 = Feedforward Neural Network with 10 hidden nodes.

For model selection, the SE (variance) should be minimised while striking a balance with the computational costs of doing so. A two-fold design had a lower variance (Table E-1), as is generally expected to be the case (e.g. Zhang & Yang, 2015). There was a law of diminishing returns for increasing numbers of iterations. A doubling in the number of iterations reduced the SE by a factor of $1/\sqrt{2} = 0.707$. To progress from 5 to 10 iterations yielded a reduction in SE of 0.07 $W \cdot kg^{-1}$, while an additional doubling in the number of iterations to 20 achieved a further decrease of 0.04 $W \cdot kg^{-1}$ to 0.12 $W \cdot kg^{-1}$. This magnitude was considered to be too small to meaningfully discriminate between models. Fifty and 100 iterations would have imposed a considerable cost without a further meaningful improvement in accuracy.

E.2 Partitioning Method for K-Fold

The partitioning of the data for k-fold cross-validation was done by participant rather than by trial throughout this thesis. It ensured that jumps from the same individual were not assigned to both the training and validation sets or to training and holdout test sets. It required code development because Matlab does not offer this facility through its *cvpartition* function. Although participant-level partitioning is a well-established procedure, it is rarely reported the benefit it brings in terms of the negative bias in the predictive error that would be avoided.

Using the same data set above (Appendix E.1), the same models were evaluated with and without participant-level partitioning. The results show that the validation loss was lower on average (0.36–0.65 $W \cdot kg^{-1}$) than the baseline case of partitioning at the participant level (Table E-5). Hence, the models underestimated the true model error by a substantial amount without partitioning at the participant level.

Table E-5. Average Loss, RMSE (W·kg^{-1}), with partitioning by trial rather than by participant, for different combination so k-folds and iterations across all 11 models. The difference with the baseline values is shown.

	Iterations					
K-Folds	**2**	**5**	**10**	**20**	**50**	**100**
2	4.01	3.91	3.90	3.88	3.86	3.85
5	3.85	3.84	3.83	3.82	3.78	3.74
10	3.66	3.66	3.66	3.66	3.64	3.61
60	3.57	3.57	3.56	3.55	3.54	3.51
Difference with Baseline:						
2	-0.36	-0.46	-0.49	-0.55	-0.61	-0.63
5	-0.65	-0.65	-0.65	-0.64	-0.63	-0.62
10	-0.61	-0.60	-0.60	-0.58	-0.58	-0.58
60	-0.55	-0.54	-0.54	-0.54	-0.51	-0.49

A related consideration was the partitioning of FPCA, where FPCs are defined strictly based on the training data without any influence from the validation data. The same models (Appendix E.1)were re-evaluated using the same method but without FPCA partitioning. Hence, the FPCs were defined using all the data (except for the holdout data). The baseline case was the original evaluation based on FPCA partitioning.

The results show that the average loss was only slightly lower than the baseline case (Table E-6), where the differences were ≤ 0.10 W·kg^{-1}. For the favoured arrangement of a two-fold design with 20 iterations, the difference was -0.04 W·kg^{-1}. All these differences were much less than the standard error. These results show that the bias introduced by allowing the FPCs to be defined using the whole data set was negligible. It suggests the FPCs themselves are quite general, differing to only a modest degree between subsets of the data.

Table E-6. Average Loss, RMSE ($W \cdot kg^{-1}$), without partitioning FPCA, for different combination so k-folds and iterations, across all 11 models. The difference with the baseline values shown in Table E-1 is also shown.

	Iterations					
K-Folds	**2**	**5**	**10**	**20**	**50**	**100**
2	4.42	4.47	4.46	4.47	4.46	4.46
5	4.45	4.45	4.44	4.42	4.37	4.32
10	4.22	4.22	4.22	4.21	4.18	4.14
60	4.08	4.08	4.07	4.06	4.01	3.95
Difference with Baseline:						
2	0.05	0.10	0.07	0.04	-0.01	-0.02
5	-0.05	-0.04	-0.04	-0.04	-0.04	-0.04
10	-0.05	-0.06	-0.04	-0.03	-0.04	-0.05
60	-0.04	-0.03	-0.03	-0.03	-0.04	-0.05

Partitioning FPCA comes with a high cost because FPCA must be run for every iteration of the cross-validation loop. For 20 iterations of two-fold cross-validation, the model function's average execution time was three times longer with FPCA partitioning than without (4143 ms vs1370 ms). Given there was no tangible benefit to FPCA partitioning for this analysis, it was removed from the accelerometer model function.

E.3 Inner Loop Partitioning

Monte Carlo cross-validation (MCCV) was a cornerstone of the methods used in this thesis. As noted in Section 5.2.4, it implied that each partition of the data was made independently of the others. Thus, for 20 iterations of cross-validation, 20 data splits were made. This approach will be called the Strict Monte Carlo strategy for the purposes of the investigation in this section. It will be compared with an alternative Monte Carlo method that can be implemented specifically with two-fold partitioning. This alternative approach involves swapping the training and validation sets for each data split to yield two evaluations and will be referred to as the Complementary strategy here.

For the Complementary strategy, the data set was split ten times, and within each iteration, two validations were performed. As such, it was a form of repeated k-fold cross validation. In the first step, one partition served as the training set, and the other served as the validation set; in the second step, the roles were reversed. In this way, $10 \times 2 = 20$ iterations of cross-validation were performed. Every observation in the data set was used for training and validation with no exceptions. On the other hand, in the Strict strategy, the data set was split 20 times, with the partitions taking only one role, either training or validation. In this method, a given observation had a 50-50 chance of being assigned to the training or validation sets. In the long run (infinite iterations), every observation would be assigned to each subset the same number of times, but with only 20 iterations, there would be unequal assignments due to chance (binomial distribution). Some observations may be used more often for training than for validation and vice versa. The Strict strategy was simpler to implement, and it could be applied universally to all k-fold designs, whereas the Complementary strategy was only applicable to the two-fold design. Therefore, the key question was whether the Complementary strategy was worth implementing for the two-fold design specifically, which played a critical role in model selection. The key metric was the validation RMSE variance, which needed to be minimised.

The Strict strategy (20 data splits to generate 20 partitions of training and validation data) was compared with the Complementary strategy (10 data splits, swapping the partitions' training and validation roles, making 20 iterations in all). All models were evaluated over 50 repetitions using the LB-CMJ_{NA} data set, following the basic data pre-processing procedure employed in Chapter 5. The means and coefficients of variation for validation RMSE were calculated over the 20 iterations and averaged across the 50 repetitions.

The results show that the Strict strategy was marginally the better one with a lower coefficient of variance overall (Table E-7). The small differences between them were approaching the typical error and may vanish with further iterations.

Table E-7. Comparison of two Monte Carlo partitioning strategies for cross-validation showing the validation RMSE. The means and coefficient of variations (CV) were calculated over 20 iterations and then averaged over 50 repetitions.

	Strict Strategy		Complementary Strategy	
Algorithms	**Mean ($W \cdot kg^{-1}$)**	**CV**	**Mean ($W \cdot kg^{-1}$)**	**CV**
LR	5.22	19.9%	5.22	19.8%
LR-RDG	5.18	18.5%	5.20	19.0%
LR-LSS	4.97	16.0%	4.98	16.0%
SVM-L	4.63	12.4%	4.63	12.3%
SVM-G	4.38	12.2%	4.36	12.3%
GPR-SE	4.66	14.0%	4.73	14.3%
GPR-M52	4.55	12.9%	4.57	12.7%
NN-5	5.31	23.4%	5.30	22.9%
NN-10	5.99	25.0%	6.08	27.9%
TR	6.65	12.3%	6.79	12.3%
TR-ENS	5.47	11.8%	5.47	11.4%
Mean	**5.18**	**16.2%**	**5.21**	**16.5%**

The Complementary strategy needed to outperform the Strict strategy, which was the simpler and universal one, to justify its inclusion as a necessary but separate approach. The results showed that it did not do so. Therefore, the Strict method was adopted for all MCCV procedures.

E.4 Bayesian Optimiser

Bayesian optimisation constructs a surrogate model based on a Gaussian process, describing the objective function's variation across one or more parameters. The optimiser updates its prior belief about the SM objective function, with each new observation. The optimiser determines the next point to sample where it believes it has the highest probability of obtaining the biggest improvement in the objective function (maximum 'expected improvement' algorithm based on the acquisition function). In

this way, the optimiser proceeds towards what it believes to be the global optimum. It is highly efficient compared to other optimisers, sampling the objective function far fewer times (Bergstra & Bengio, 2012). It is therefore suited towards optimising computationally expensive objective functions, such as the AM in this thesis, which includes data pre-processing and the model fitting itself. It can also handle the AM's stochastic behaviour that arises from the random sub-sampling of cross-validation. In summary, the Bayesian optimiser appeared to be the ideal candidate for the optimiser in this thesis.

However, it became apparent in preliminary studies that the Bayesian optimiser did not provide consistent answers. This inconsistency can be illustrated using the GPR model from Chapter 5, which has three categorical hyperparameters, allowing optimal models to be conveniently grouped together. It also had one numerical hyperparameter, but that is irrelevant for this investigation. The Bayesian optimiser was run 50 times using the same data set, completing 50 iterations each time before reporting a final optimal model. The optimal model was based on the feasible minimum determined by the surrogate model rather than the observed minimum because the former was more representative of the model.

The frequencies of the optimal models that were produced are shown in Figure E-1, which has grouped them by basis function, kernel function and the standardisation option. There would be one optimal model in theory, but the results show a wide range of apparently optimal models. The new optimisation procedure introduced in this thesis also produced more than one optimal model (Section 5.3.2), but the range of possible models generated by multiple runs of the Bayesian optimiser was much more extensive and correspondingly more difficult to resolve. One approach to resolve this was selecting the model that appeared most frequently (Matern 3/2 or 5/2 kernel, linear basis function with standardisation). However, further analysis involving more observations revealed that there were other models with a lower predictive error. With its efficient but parsimonious approach for selecting where to take the next observations, the Bayesian optimiser could only ever obtain a partial view of the parameter space. It determined where the most likely minimum lay based on the observations gathered up to that point based on maximum likelihood. That did not

mean that the true minimum did lie there, only that that location was the most likely. With further observations, the model could well have reached another conclusion. Indeed, the level of noise (variance in the CV estimates) confounded this situation further.

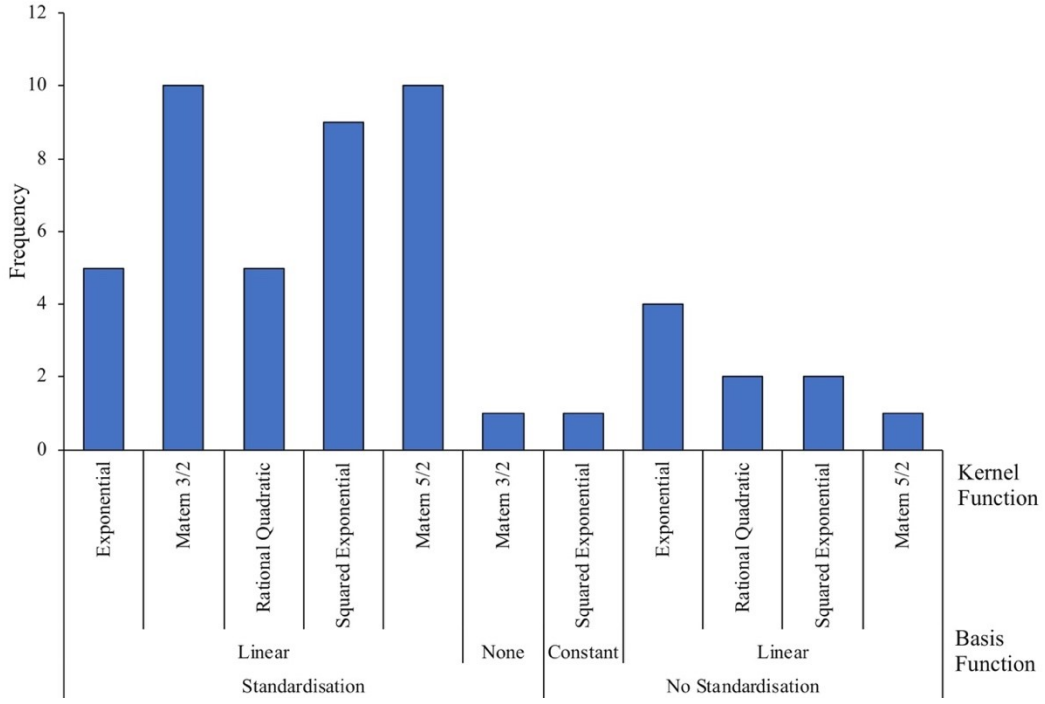


Figure E-1. Different optimal models from the same data set for the Chapter 5 GPR model. Over 50 runs, the Bayesian optimiser produced a range of optimal models with different kernel functions, basis functions and standardisation options.

The optimal models themselves proved to be at odds with the results from the surveys made of the parameter space, such as in Chapter 5. Most of the models included the standardisation option (Figure E-1, left half of the plot), but all the final optimal GPR models in Chapter 5 did not use standardisation (Figure 5-12L). Models based on a linear basis function with standardisation appeared more frequently, and among those, the Matern 3/2 and 5/2 kernels were the most common, just ahead of the squared exponential kernel (Figure E-1). This, too, contradicted the results from Chapter 5 (Figure 5-12I), where models with a linear basis function tended to have a higher validation error than those with a constant basis function or no basis at all.

Furthermore, the Bayesian optimiser selected the exponential kernel on nine occasions (Figure E-1), but the survey from Chapter 5 revealed that models using that kernel were the least accurate (Figure 5-12J). It was therefore concluded that given these and many other contradictory results, the Bayesian optimiser should not be employed for the research in this thesis.

The Bayesian optimiser had only a partial view of the parameter space, based on few AM observations, which were themselves noisy due to data sub-sampling variance. Without sufficient observations to average out such variance, the Bayesian optimiser could get a false impression of the true AM behaviour, which may account for the inconsistent results. More iterations of the optimiser did not improve the situation as instability in the optimal parameter values persisted. There was also evidence from the SM progression series that stabilised parameters could begin to vary erratically after 200-300 iterations. The steep rise in the acquisition function overhead also made running the Bayesian optimiser beyond 500 iterations increasingly impractical, as the next section shows.

E.5 Bayesian Optimiser Overhead

The Bayesian optimiser relies on its acquisition function to determine the next point to sample in the search. According to the expected-improvement algorithm, the acquisition function itself must be maximised, which can be a computationally intensive task. As discussed in Chapter 5, various methods have been proposed to make the algorithm more efficient, including Monte Carlo integration (Wilson et al., 2018) and batch processing (Snoek et al., 2012; Taddy et al., 2009), which is incorporated into the Matlab implementation. Maximising the acquisition function makes up a substantial proportion of the Bayesian overhead, the other being the fitting of the surrogate model. When running the Bayesian optimiser in Matlab, the rising cost of this overhead becomes noticeable quite early on (within 30 iterations). Attempts to improve the optimisation's reliability initially centred on increasing the number of iterations, but the computational costs became prohibitive, as this section will illustrate.

The Bayesian optimiser was run on the GPR model from Chapter 5 for 1000 iterations, a more extensive search than might usually be performed. In comparison, the Matlab default is 30 iterations, which are usually sufficient for many functions. The optimiser logged the objective function's (AM's) execution and overall iteration times. The overhead at each step was defined as the overall iteration time minus the objective function execution time. The overhead ratio for each iteration was defined as the overhead time divided by the objective function execution time, averaged across all iterations (1.653 s).

The results show that the accumulated time for overall optimiser execution rose exponentially (Figure E-2A). It required 7 mins 50 s to complete 100 iterations, 20 mins 47 s for 200 iterations, 1 hr 40 mins for 500 iterations, almost 4 hours for 750 iterations and a little over 9 hours for 1000 iterations. The processing was performed on an iMac 2017, MacOS 10.14.6, 3 GHz Intel Core i5 with a Matlab R2020a ODE benchmark of 0.4620 s). A good approximation for the accumulated time in the latter stages, T, based on the final 200 iterations, was given by $T \sim 5\times10^{-5} i^{2.92}$, where i represented the iteration count. Extrapolating to 2000 iterations using this formula, the Bayesian optimiser would require nearly five days of processing to achieve the same level of coverage in Chapter 5. In contrast, the new optimisation procedure required around 3 hours.

The overhead ratio also rose exponentially but at a more consistent rate (Figure E-2B). Initially, the optimiser overhead was a factor of two, that is, requiring twice the processing cost of the accelerometer model. After 200 iterations, the overhead ratio was 5.5 times and around 430 iterations, it had past a factor 10. By 1000 iterations, the overhead ratio had exceeded 60. From these figures, it was clear that the optimiser overhead had become the overwhelming factor.

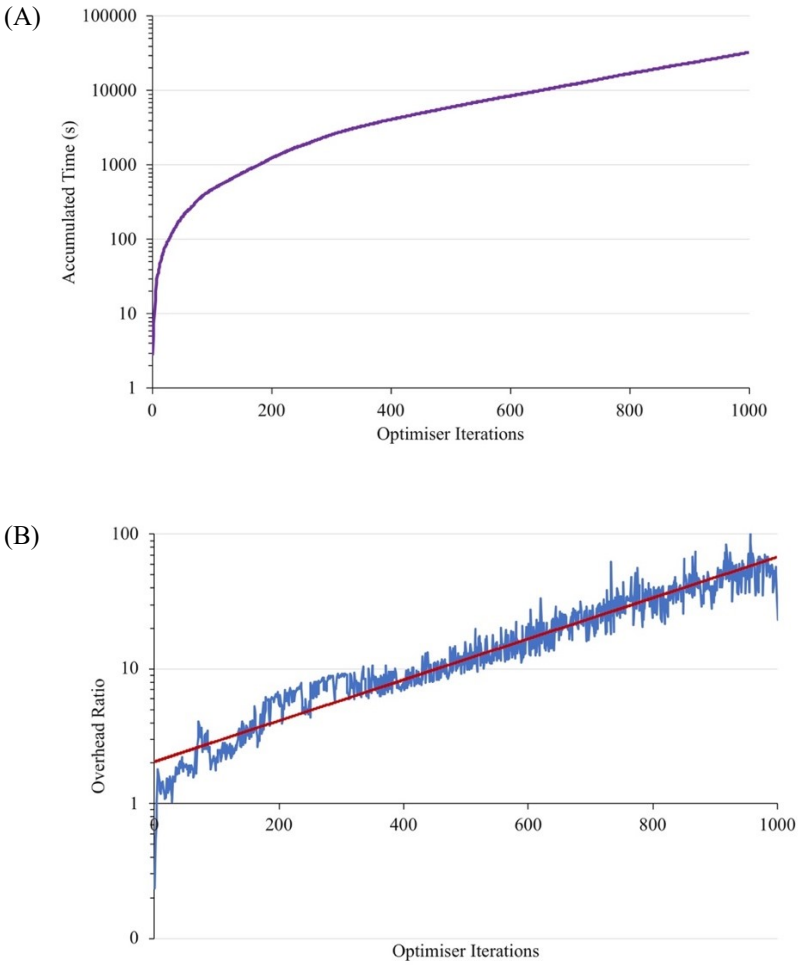


Figure E-2. Bayesian Optimisation for the Chapter 5 GPR model. (A) Accumulated time for the optimisation (1000 s ~ 16 mins; 10000 s ~ 2¾ hrs; 100000 s ~ 28 hrs); (B) Overhead ratio (optimiser overhead / objective function execution times) showing exponential growth with a best fit straight line.

In the Matlab implementation, the batch processing technique (block coordinate descent, BCD) was activated after 300 iterations to make the process more efficient, which explained the slight flattening of the exponential growth rate (Figure E-2A). Even with this improved efficiency, it was concluded that given the approximately cubic geometric rise in processing time, it would not be practical to run the Bayesian optimiser for the number of iterations required to locate the optimum with consistency.

E.6 SM Noise related to MCCV Iterations

The AM loss has stochastic behaviour as a result of the sub-sampling variance of the cross validation procedure (Section5.2.5). Hence, the AM appeared noisy to an optimiser which would be uncertain about the true value of the AM loss for a given set of parameter values. As a consequence, it is more difficult for the optimiser to locate the global minimum. This task can be made easier by increasing the number of MCCV iterations, which reduces the SE in the loss estimate and hence, the apparent noise. It follows that the SM noise will reflect the AM noise level (SE in the loss estimate) provided sufficient numbers of observations have been made. The SM noise level sets a lower limit for the confidence interval of the SM predictions (Figure 5-10, second row), and so it may be expected that reducing AM noise should lessen SM noise and therefore narrow the SM confidence interval. Thus, with less uncertainty, the SM becomes more accurate, more quickly, which should make the optimisation more efficient. However, increasing the number of MCCV iterations to achieve this comes at a higher cost for the AM, which would reduce the total number of observations that can be made for the same computational budget. It therefore needed to be established if increasing the number of MCCV iterations could make a worthwhile improvement in the accuracy of the SM model.

The Chapter 6 GPR model was chosen for the investigation, which formed part of a line of research that looked into ways of improving the efficacy of the optimisation procedure. The performance of the optimisation procedure needed to be improved for models with numerical parameters in particular. The Chapter 6 model, with its four numerical parameters covering the time window and smoothing levels, was a suitable test case. The optimisation procedure was repeated for 2, 10, 20 and 100 MCCV iterations to determine whether increasing the number of iterations in the cross-validation loop would reduce the SM noise level. A two-fold cross validation design was adopted, and the random search gathered 2000 observations with the SM updated every 50 iterations. No progressive restriction was imposed on the random search, the technique that was adopted in Chapters 6 onwards. Hence, as in Chapter 5, the observations were approximately evenly spread throughout the search space. This approach ensured the benefits of progressively focusing the search on a narrower area,

thereby increasing the density of observations, which did not influence the results. The SM fitted noise was averaged over the final 20 models out of the 100 recorded, the period where it had stabilised.

The results show that the SM noise fell progressively with increasing numbers of MCCV iterations, as expected, but there was a law of diminishing returns (Table E-8). The noise reduction was substantial when moving from 2 to 10 iterations, and there was further improvement with 20 iterations (0.218 $W \cdot kg^{-1}$). However, going to 100 iterations only brought a marginal improvement (0.029 $W \cdot kg^{-1}$) for a considerable additional cost. The conclusion was that increasing the number of MCCV iterations was not the best way of improving the efficacy of the optimisation procedure. The best that might be achieved, represented by the case of 100 iterations, did not deliver a material reduction in noise. The additional overhead was not considered to be worth it.

Table E-8. Surrogate Model noise level for different numbers of cross-validation iterations when applied to the Chapter 6 model.

MCCV Iterations	SM Noise ($W \cdot kg^{-1}$)
2	0.417
10	0.243
20	0.218
100	0.189

REFERENCES

Ache-Dias, J., Dal Pupo, J., Gheller, R. G., Külkamp, W., & Moro, A. R. P. (2016). Power Output Prediction From Jump Height and Body Mass Does Not Appropriately Categorize or Rank Athletes: *Journal of Strength and Conditioning Research*, *30*(3), 818–824. https://doi.org/10.1519/JSC.0000000000001150

Adamson, G. T., & Whitney, R. J. (1971). Critical Appraisal of Jumping as a Measure of Human Power. In J. Vredenbregt & J. Wartenweiler (Eds.), *Medicine and Sport Science* (Vol. 6, pp. 208–211). S. Karger AG. https://doi.org/10.1159/000392173

Aggarwal, K., Kirchmeyer, M., Yadav, P., Keerthi, S. S., & Gallinari, P. (2020). Benchmarking Regression Methods: A comparison with CGAN. *ArXiv:1905.12868 [Cs, Stat]*. http://arxiv.org/abs/1905.12868

Allen, D. M. (1974). The Relationship Between Variable Selection and Data Agumentation and a Method for Prediction. *Technometrics*, *16*(1), 125–127. https://doi.org/10.1080/00401706.1974.10489157

Ambroise, C., & McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences*, *99*(10), 6562–6566. https://doi.org/10.1073/pnas.102102699

Amonette, W. E., Brown, L. E., De Witt, J. K., Dupler, T. L., Tran, T. T., Tufano, J. J., & Spiering, B. A. (2012). Peak Vertical Jump Power Estimations in Youths and Young Adults: *Journal of Strength and Conditioning Research*, *26*(7), 1749–1755. https://doi.org/10.1519/JSC.0b013e3182576f1e

Ancillao, A., Tedesco, S., Barton, J., & O'Flynn, B. (2018). Indirect Measurement of Ground Reaction Forces and Moments by Means of Wearable Inertial Sensors: A Systematic Review. *Sensors*, *18*(8), 2564. https://doi.org/10.3390/s18082564

Andén, J., & Mallat, S. (2014). Deep Scattering Spectrum. *IEEE Transactions on Signal Processing*, *62*(16), 4114–4128. https://doi.org/10.1109/TSP.2014.2326991

Andersson, H., Raastad, T., Nilsson, J., Paulsen, G., Garthe, I., & Kadi, F. (2008). Neuromuscular Fatigue and Recovery in Elite Female Soccer: Effects of Active Recovery. *Medicine & Science in Sports & Exercise*, *40*(2), 372–380. https://doi.org/10.1249/mss.0b013e31815b8497

Anscombe, F. J. (1967). Topics in the Investigation of Linear Relations Fitted by the Method of Least Squares. *Journal of the Royal Statistical Society: Series B (Methodological)*, *29*(1), 1–29. https://doi.org/10.1111/j.2517-6161.1967.tb00672.x

Apostolidis, N., Nassis, G. P., Bolatoglou, T., & Geladas, N. D. (2004). Physiological and technical characteristics of elite young basketball players. *The Journal of Sports Medicine and Physical Fitness*, *44*(2), 157–163.

Aragón, L. F. (2000). Evaluation of Four Vertical Jump Tests: Methodology, Reliability, Validity, and Accuracy. *Measurement in Physical Education and Exercise Science*, *4*(4), 215–228. https://doi.org/10.1207/S15327841MPEE0404_2

Aragón-Vargas, L. F., & Gross, M. M. (1997). Kinesiological Factors in Vertical Jump Performance: Differences among Individuals. *Journal of Applied Biomechanics*, *13*(1), 24–44. https://doi.org/10.1123/jab.13.1.24

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*(0), 40–79. https://doi.org/10.1214/09-SS054

Arnason, A., Sigurdsson, S. B., Gudmundsson, A., Holme, I., Engebretsen, L., & Bahr, R. (2004). Physical Fitness, Injuries, and Team Performance in Soccer: *Medicine & Science in Sports & Exercise*, *36*(2), 278–285. https://doi.org/10.1249/01.MSS.0000113478.92945.CA

Arteaga, R., Dorado, C., Chavarren, J., & Calbet, J. A. (2000). Reliability of jumping performance in active men and women under different stretch loading conditions. *The Journal of Sports Medicine and Physical Fitness*, *40*(1), 26–34.

Atkinson, G., & Nevill, A. M. (1998). Statistical Methods For Assessing Measurement Error (Reliability) in Variables Relevant to Sports Medicine: *Sports Medicine*, *26*(4), 217–238. https://doi.org/10.2165/00007256-199826040-00002

Attal, F., Mohammed, S., Dedabrishvili, M., Chamroukhi, F., Oukhellou, L., & Amirat, Y. (2015). Physical Human Activity Recognition Using Wearable Sensors. *Sensors*, *15*(12), 31314–31338. https://doi.org/10.3390/s151229858

Bachmann, E. R., Xiaoping Yun, & Peterson, C. W. (2004). An investigation of the effects of magnetic variations on inertial/magnetic orientation sensors. *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004*, 1115-1122 Vol.2. https://doi.org/10.1109/ROBOT.2004.1307974

Bair, E., Hastie, T., Paul, D., & Tibshirani, R. (2006). Prediction by Supervised Principal Components. *Journal of the American Statistical Association*, *101*(473), 119–137. https://doi.org/10.1198/016214505000000628

Baker, D. (2002). Differences in strength and power among junior-high, senior-high, college-aged, and elite professional rugby league players. *Journal of Strength and Conditioning Research*, *16*(4), 581–585.

Baker, D., & Nance, S. (1999a). The Relation Between Running Speed and Measures of Strength and Power in Professional Rugby League Players. *Journal of Strength & Conditioning Research*, *13*(3), 230–235.

Baker, D., & Nance, S. (1999b). The Relation Between Strength and Power in Professional Rugby League Players. *Journal of Strength & Conditioning Research*, *13*(3), 224–229.

Barnett, A. (2006). Using recovery modalities between training sessions in elite athletes: Does it help? *Sports Medicine (Auckland, N.Z.)*, *36*(9), 781–796. https://doi.org/10.2165/00007256-200636090-00005

Bartlett, R., Wheat, J., & Robins, M. (2007). Is movement variability important for sports biomechanists? *Sports Biomechanics*, *6*(2), 224–243. https://doi.org/10.1080/14763140701322994

Bates, B. T., Dufek, J. S., & Davis, H. P. (1992). The effect of trial size on statistical power. *Medicine and Science in Sports and Exercise*, *24*(9), 1059–1065.

Baumann, D., & Baumann, K. (2014). Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation. *Journal of Cheminformatics*, *6*(1), 47. https://doi.org/10.1186/s13321-014-0047-1

Baumann, K. (2003). Cross-validation as the objective function for variable-selection techniques. *TrAC Trends in Analytical Chemistry*, *22*(6), 395–406. https://doi.org/10.1016/S0165-9936(03)00607-1

Beek, P. J., Peper, C. E., & Stegeman, D. F. (1995). Dynamical models of movement coordination. *Human Movement Science*, *14*(4–5), 573–608. https://doi.org/10.1016/0167-9457(95)00028-5

Bergamini, E., Ligorio, G., Summa, A., Vannozzi, G., Cappozzo, A., & Sabatini, A. M. (2014). Estimating Orientation Using Magnetic and Inertial Sensors and Different Sensor Fusion Approaches: Accuracy Assessment in Manual and Locomotion Tasks. *Sensors*, *14*(10), 18625–18649. https://doi.org/10.3390/s141018625

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *J. Mach. Learn. Res.*, *13*(null), 281–305.

Bischl, B., Mersmann, O., Trautmann, H., & Weihs, C. (2012). Resampling Methods for Meta-Model Validation with Recommendations for Evolutionary Computation. *Evolutionary Computation*, *20*(2), 249–275. https://doi.org/10.1162/EVCO_a_00069

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Blank, P., Hoßbach, J., Schuldhaus, D., & Eskofier, B. M. (2015). Sensor-based stroke detection and stroke type classification in table tennis. *Proceedings of the 2015 ACM International Symposium on Wearable Computers - ISWC '15*, 93–100. https://doi.org/10.1145/2802083.2802087

Bobbert, M. F., Casius, L. J. R., Sijpkens, I. W. T., & Jaspers, R. T. (2008). Humans adjust control to initial squat depth in vertical squat jumping. *Journal of Applied Physiology*, *105*(5), 1428–1440. https://doi.org/10.1152/japplphysiol.90571.2008

Bobbert, M. F., Huijing, P. A., & van Ingen Schenau, G. J. (1986). A model of the human triceps surae muscle-tendon complex applied to jumping. *Journal of Biomechanics*, *19*(11), 887–898. https://doi.org/10.1016/0021-9290(86)90184-3

Bobbert, M. F., Huijing, P. A., & van Ingen Schenau, G. J. (1987a). Drop jumping. I. The influence of jumping technique on the biomechanics of jumping. *Medicine and Science in Sports and Exercise*, *19*(4), 332–338.

Bobbert, M. F., Huijing, P. A., & van Ingen Schenau, G. J. (1987b). Drop jumping. II. The influence of dropping height on the biomechanics of drop jumping. *Medicine and Science in Sports and Exercise*, *19*(4), 339–346.

Bobbert, M. F., & van Ingen Schenau, G. J. (1988). Coordination in vertical jumping. *Journal of Biomechanics*, *21*(3), 249–262. https://doi.org/10.1016/0021-9290(88)90175-3

Bobbert, M. F., & Van Zandwijk, J. P. (1999). Dynamics of force and muscle stimulation in human vertical jumping: *Medicine & Science in Sports & Exercise*, *31*(2), 303–310. https://doi.org/10.1097/00005768-199902000-00015

Borràs, X., Balius, X., Drobnic, F., & Galilea, P. (2011). Vertical Jump Assessment on Volleyball: A Follow-Up of Three Seasons of a High-Level Volleyball Team: *Journal of Strength and Conditioning Research*, *25*(6), 1686–1694. https://doi.org/10.1519/JSC.0b013e3181db9f2e

Bosco, C., Luhtanen, P., & Komi, P. V. (1983). A simple method for measurement of mechanical power in jumping. *European Journal of Applied Physiology and Occupational Physiology*, *50*(2), 273–282. https://doi.org/10.1007/BF00422166

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory - COLT '92*, 144–152. https://doi.org/10.1145/130385.130401

Bouktif, S., Fiaz, A., Ouni, A., & Serhani, M. (2018). Optimal Deep Learning LSTM Model for Electric Load Forecasting using Feature Selection and Genetic Algorithm: Comparison with Machine Learning Approaches †. *Energies*, *11*(7), 1636. https://doi.org/10.3390/en11071636

Boulgouris, N. V., Plataniotis, K. N., & Hatzinakos, D. (2004). Gait recognition using dynamic time warping. *IEEE 6th Workshop on Multimedia Signal Processing, 2004.*, 263–266. https://doi.org/10.1109/MMSP.2004.1436543

Braga-Neto, U. M., & Dougherty, E. R. (2004). Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, *20*(3), 374–380. https://doi.org/10.1093/bioinformatics/btg419

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/BF00058655

Breiman, L., & Spector, P. (1992). Submodel Selection and Evaluation in Regression. The X-Random Case. *International Statistical Review / Revue Internationale de Statistique*, *60*(3), 291. https://doi.org/10.2307/1403680

Brochu, E., Cora, V. M., & de Freitas, N. (2010). A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. *ArXiv.Org*, 49. https://arxiv.org/abs/1012.2599

Bruton, M. R., O'Dwyer, N., & Adams, R. (2013). Sex differences in the kinematics and neuromuscular control of landing: Biological, environmental and sociocultural factors. *Journal of Electromyography and Kinesiology*, *23*(4), 747–758. https://doi.org/10.1016/j.jelekin.2013.04.012

Buckthorpe, M., Morris, J., & Folland, J. P. (2012). Validity of vertical jump measurement devices. *Journal of Sports Sciences*, *30*(1), 63–69. https://doi.org/10.1080/02640414.2011.624539

Bull, A. D. (2011). Convergence rates of efficient global optimization algorithms. *ArXiv:1101.3501 [Math, Stat]*. http://arxiv.org/abs/1101.3501

Burdack, J., Horst, F., Giesselbach, S., Hassan, I., Daffner, S., & Schöllhorn, W. I. (2019). Systematic Comparison of the Influence of Different Data Preprocessing Methods on the Classification of Gait Using Machine Learning. *ArXiv:1911.04335 [Cs, Stat]*. http://arxiv.org/abs/1911.04335

Burman, P. (1989). A Comparative Study of Ordinary Cross-Validation, v-Fold Cross-Validation and the Repeated Learning-Testing Methods. *Biometrika*, *76*(3), 503. https://doi.org/10.2307/2336116

Button, C., MacLeod, M., Sanders, R., & Coleman, S. (2003). Examining movement variability in the basketball free-throw action at different skill levels. *Research Quarterly for Exercise and Sport*, *74*(3), 257–269. https://doi.org/10.1080/02701367.2003.10609090

Canavan, P. K., & Vescovi, J. D. (2004). Evaluation of Power Prediction Equations: Peak Vertical Jumping Power in Women: *Medicine & Science in Sports & Exercise*, *36*(9), 1589–1593. https://doi.org/10.1249/01.MSS.0000139802.96395.AC

Casartelli, N., Müller, R., & Maffiuletti, N. A. (2010). Validity and Reliability of the Myotest Accelerometric System for the Assessment of Vertical Jump Height: *Journal of Strength and Conditioning Research*, *24*(11), 3186–3193. https://doi.org/10.1519/JSC.0b013e3181d8595c

Castagna, C., & Castellini, E. (2013). Vertical Jump Performance in Italian Male and Female National Team Soccer Players: *Journal of Strength and Conditioning Research*, *27*(4), 1156–1161. https://doi.org/10.1519/JSC.0b013e3182610999

Castagna, C., Ganzetti, M., Ditroilo, M., Giovannelli, M., Rocchetti, A., & Manzi, V. (2013). Concurrent Validity of Vertical Jump Performance Assessment Systems: *Journal of Strength and Conditioning Research*, *27*(3), 761–768. https://doi.org/10.1519/JSC.0b013e31825dbcc5

Cavanagh, P. R. (1987). The Biomechanics of Lower Extremity Action in Distance Running. *Foot & Ankle*, *7*(4), 197–217. https://doi.org/10.1177/107110078700700402

Cawley, G. C., & Talbot, N. L. C. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research*, *11*, 2079–2107.

Chau, T., Young, S., & Redekop, S. (2005). Managing variability in the summary and comparison of gait data. *Journal of NeuroEngineering and Rehabilitation*, 20.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chelly, S. M., & Denis, C. (2001). Leg power and hopping stiffness: Relationship with sprint running performance: *Medicine and Science in Sports and Exercise*, 326–333. https://doi.org/10.1097/00005768-200102000-00024

Chen, L., Li, J., Zhang, Y., Feng, K., Wang, S., Zhang, Y., Huang, T., Kong, X., & Cai, Y. (2018). Identification of gene expression signatures across different

types of neural stem cells with the Monte-Carlo feature selection method. *Journal of Cellular Biochemistry*, *119*(4), 3394–3403. https://doi.org/10.1002/jcb.26507

Cheng, Y. (2019). Semi-supervised Learning for Neural Machine Translation. In Y. Cheng, *Joint Training for Neural Machine Translation* (pp. 25–40). Springer Singapore. https://doi.org/10.1007/978-981-32-9748-7_3

Choukou, M.-A., Laffaye, G., & Taiar, R. (2014). Reliability and validity of an accele-rometric system for assessing vertical jumping performance. *Biology of Sport*, *31*(1), 55–62. https://doi.org/10.5604/20831862.1086733

Ciacci, S., Merni, F., Bartolomei, S., & Di Michele, R. (2017). Sprint start kinematics during competition in elite and world-class male and female sprinters. *Journal of Sports Sciences*, *35*(13), 1270–1278. https://doi.org/10.1080/02640414.2016.1221519

Claudino, J. G., Cronin, J., Mezêncio, B., McMaster, D. T., McGuigan, M., Tricoli, V., Amadio, A. C., & Serrão, J. C. (2017). The countermovement jump to monitor neuromuscular status: A meta-analysis. *Journal of Science and Medicine in Sport*, *20*(4), 397–402. https://doi.org/10.1016/j.jsams.2016.08.011

Cockcroft, J., Robyn, A., Louw, Q., & Bayne, H. (2019). *A large-scale numerical analysis of unimodal and bimodal features in force plate data measured during vertical countermovement jumping*. *37*(1), Article 102.

Coffey, N., Harrison, A. J., Donoghue, O. A., & Hayes, K. (2011). Common functional principal components analysis: A new approach to analyzing human movement data. *Human Movement Science*, *30*(6), 1144–1166. https://doi.org/10.1016/j.humov.2010.11.005

Cohen, J. (1992). A power primer. *Quantitative Methods in Psychology*, *112*(1), 155–159.

Cooper, G., Sheret, I., McMillian, L., Siliverdis, K., Sha, N., Hodgins, D., Kenney, L., & Howard, D. (2009). Inertial sensor-based knee flexion/extension angle estimation. *Journal of Biomechanics*, *42*(16), 2678–2685. https://doi.org/10.1016/j.jbiomech.2009.08.004

Cormack, S. J., Newton, R. U., & McGuigan, M. R. (2008). Neuromuscular and Endocrine Responses of Elite Players to an Australian Rules Football Match. *International Journal of Sports Physiology and Performance*, *3*(3), 359–374. https://doi.org/10.1123/ijspp.3.3.359

Cormack, S. J., Newton, R. U., McGuigan, M. R., & Doyle, T. L. A. (2008). Reliability of Measures Obtained During Single and Repeated Countermovement Jumps. *International Journal of Sports Physiology and Performance*, *3*(2), 131–144. https://doi.org/10.1123/ijspp.3.2.131

Cormie, P., McGuigan, M. R., & Newton, R. U. (2011). Developing Maximal Neuromuscular Power. *Sports Med*, 22.

Crane, E. A., Cassidy, R. B., Rothman, E. D., & Gerstner, G. E. (2010). Effect of registration on cyclical kinematic data. *Journal of Biomechanics*, *43*(12), 2444–2447. https://doi.org/10.1016/j.jbiomech.2010.04.024

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative Adversarial Networks: An Overview. *IEEE Signal Processing Magazine*, *35*(1), 53–65. https://doi.org/10.1109/MSP.2017.2765202

Crewther, B. T., Kilduff, L. P., Cunningham, D. J., Cook, C., Owen, N., & Yang, G.-Z. (2011). Validating Two Systems for Estimating Force and Power. *International Journal of Sports Medicine*, *32*(04), 254–258. https://doi.org/10.1055/s-0030-1270487

Cronin, J., & Hansen, K. T. (2005). Strength and power predictors of sports speed. *Journal of Strength and Conditioning Research*, *19*(2), 349–357. https://doi.org/10.1519/14323.1

Cronin, J., & Sleivert, G. (2005). Challenges in Understanding the Influence of Maximal Power Training on Improving Athletic Performance: *Sports Medicine*, *35*(3), 213–234. https://doi.org/10.2165/00007256-200535030-00003

Cuspinera, L. P., Uetsuji, S., Morales, F. J. O., & Roggen, D. (2016). Beach volleyball serve type recognition. *Proceedings of the 2016 ACM International Symposium on Wearable Computers - ISWC '16*, 44–45. https://doi.org/10.1145/2971763.2971781

Cust, E. E., Sweeting, A. J., Ball, K., & Robertson, S. (2019). Machine and deep learning for sport-specific movement recognition: A systematic review of model development and performance. *Journal of Sports Sciences*, *37*(5), 568–600. https://doi.org/10.1080/02640414.2018.1521769

Davids, K., Glazier, P., Ara??jo, D., & Bartlett, R. (2003). Movement Systems as Dynamical Systems: The Functional Role of Variability and its Implications for Sports Medicine. *Sports Medicine*, *33*(4), 245–260. https://doi.org/10.2165/00007256-200333040-00001

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press.

De Ruiter, C. J., Van Leeuwen, D., Heijblom, A., Bobbert, M. F., & De Haan, A. (2006). Fast Unilateral Isometric Knee Extension Torque Development and Bilateral Jump Height: *Medicine & Science in Sports & Exercise*, *38*(10), 1843–1852. https://doi.org/10.1249/01.mss.0000227644.14102.50

de Vries, W. H. K., Veeger, H. E. J., Baten, C. T. M., & van der Helm, F. C. T. (2009). Magnetic distortion in motion labs, implications for validating inertial magnetic sensors. *Gait & Posture*, *29*(4), 535–541. https://doi.org/10.1016/j.gaitpost.2008.12.004

Deluzio, K. J., & Astephen, J. L. (2007). Biomechanical features of gait waveform data associated with knee osteoarthritis. *Gait & Posture*, *25*(1), 86–93. https://doi.org/10.1016/j.gaitpost.2006.01.007

Derie, R., Robberechts, P., Van den Berghe, P., Gerlo, J., De Clercq, D., Segers, V., & Davis, J. (2020). Tibial Acceleration-Based Prediction of Maximal Vertical Loading Rate During Overground Running: A Machine Learning Approach. *Frontiers in Bioengineering and Biotechnology*, *8*, 33. https://doi.org/10.3389/fbioe.2020.00033

Detanico, D., Dal Pupo, J., Graup, S., & dos Santos, S. G. (2016). Vertical jump performance and isokinetic torque discriminate advanced and novice judo athletes. *Kinesiology*, *48*(2), 223–228. https://doi.org/10.26582/k.48.2.8

Dietterich, T. G. (2000a). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, *40*(2), 139–157. https://doi.org/10.1023/A:1007607513941

Dietterich, T. G. (2000b). Ensemble Methods in Machine Learning. In G. Goos, J. Hartmanis, & J. van Leeuwen (Eds.), *Multiple Classifier Systems* (Vol. 1857, pp. 1–15). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45014-9_1

Domhan, T., Springenberg, J. T., & Hutter, F. (2015). Speeding up Automatic Hyperparameter Optimization of Deep Neural Networks by Extrapolation of Learning Curves. *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, 3460–3468.

Domire, Z. J., & Challis, J. H. (2007). The influence of squat depth on maximal vertical jump performance. *Journal of Sports Sciences*, *25*(2), 193–200. https://doi.org/10.1080/02640410600630647

Donà, G., Preatoni, E., Cobelli, C., Rodano, R., & Harrison, A. J. (2009). Application of functional principal component analysis in race walking: An emerging methodology. *Sports Biomechanics*, *8*(4), 284–301. https://doi.org/10.1080/14763140903414425

Donoghue, O. A., Harrison, A. J., Coffey, N., & Hayes, K. (2008). Functional Data Analysis of Running Kinematics in Chronic Achilles Tendon Injury: *Medicine & Science in Sports & Exercise*, *40*(7), 1323–1335. https://doi.org/10.1249/MSS.0b013e31816c4807

Dowling, J. J., & Vamos, L. (1993). Identification of Kinetic and Temporal Factors Related to Vertical Jump Performance. *Journal of Applied Biomechanics*, *9*(2), 95–110. https://doi.org/10.1123/jab.9.2.95

Draminski, M., Rada-Iglesias, A., Enroth, S., Wadelius, C., Koronacki, J., & Komorowski, J. (2008). Monte Carlo feature selection for supervised classification. *Bioinformatics*, *24*(1), 110–117. https://doi.org/10.1093/bioinformatics/btm486

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1996). Support Vector Regression Machines. *Proceedings of the 9th International Conference on Neural Information Processing Systems*, 155–161.

Dufek, J. S., Bates, B. T., & Davis, H. P. (1995). The effect of trial size and variability on statistical power: *Medicine & Science in Sports & Exercise*, *27*(2), 288–295. https://doi.org/10.1249/00005768-199502000-00021

Efron, B. (1983). Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *Journal of the American Statistical Association*, *78*(382), 316–331. https://doi.org/10.1080/01621459.1983.10477973

Efron, B. (1986). How Biased is the Apparent Error Rate of a Prediction Rule? *Journal of the American Statistical Association*, 11.

Efron, B., & Tibshirani, R. (1997). Improvements on Cross-Validation: The .632+ Bootstrap Method. *Journal of the American Statistical Association*, *92*(438), 548. https://doi.org/10.2307/2965703

Eggensperger, K., Feurer, M., Hutter, F., Bergstra, J., Snoek, J., Hoos, H. H., & Leyton-brown, K. (2013). Towards an empirical foundation for assessing Bayesian optimization of hyperparameters. *In NIPS Workshop on Bayesian Optimization in Theory and Practice*.

Elvin, N. G., Elvin, A. A., & Arnoczky, S. P. (2007). Correlation between Ground Reaction Force and Tibial Acceleration in Vertical Jumping. *Journal of Applied Biomechanics*, *23*(3), 180–189. https://doi.org/10.1123/jab.23.3.180

Engels, R., & Theusinger, C. (1998). Using a Data Metric for Preprocessing Advice for Data Mining Applications. *In Proceedings of the European Conference on Artificial Intelligence (ECAI-98*, 430–434.

Epifanio, I., Ávila, C., Page, Á., & Atienza, C. (2008). Analysis of multiple waveforms by means of functional principal component analysis: Normal versus pathological patterns in sit-to-stand movement. *Medical & Biological Engineering & Computing*, *46*(6), 551–561. https://doi.org/10.1007/s11517-008-0339-6

Escalante, H. J., Montes, M., & Sucar, E. (2010). Ensemble particle swarm model selection. *The 2010 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN.2010.5596915

Favre, J., Jolles, B. M., Aissaoui, R., & Aminian, K. (2008). Ambulatory measurement of 3D knee joint angle. *Journal of Biomechanics*, *41*(5), 1029–1035. https://doi.org/10.1016/j.jbiomech.2007.12.003

Feltner, M. E., Bishop, E. J., & Perez, C. M. (2004). Segmental and Kinetic Contributions in Vertical Jumps Performed with and without an Arm Swing. *Research Quarterly for Exercise and Sport*, *75*(3), 216–230. https://doi.org/10.1080/02701367.2004.10609155

Ferber, D. R. (2006). A biomechanical perspective of predicting injury risk in running. *International SportMed Journal*, 12.

Feurer, M., & Hutter, F. (2019). Hyperparameter Optimization. In F. Hutter, L. Kotthoff, & J. Vanschoren (Eds.), *Automated Machine Learning* (pp. 3–33). Springer International Publishing. https://doi.org/10.1007/978-3-030-05318-5_1

Feynman, R. P. (1964). *The Law of Gravitation, an Example of Physical Law: Vol. Messenger Lectures*. https://youtu.be/j3mhkYbznBk

Field, A. P. (2013). *Discovering statistics using IBM SPSS statistics: And sex and drugs and rock 'n' roll* (4th edition). Sage.

Floria, P., Gómez-Landero, L. A., & Harrison, A. J. (2014). Variability in the Application of Force During the Vertical Jump in Children and Adults. *Journal of Applied Biomechanics*, *30*(6), 679–684. https://doi.org/10.1123/jab.2014-0043

Forner-Cordero, A., Mateu-Arce, M., Forner-Cordero, I., Alcántara, E., Moreno, J. C., & Pons, J. L. (2008). Study of the motion artefacts of skin-mounted

inertial sensors under different attachment conditions. *Physiological Measurement*, *29*(4), N21–N31. https://doi.org/10.1088/0967-3334/29/4/N01

Forrester, S. E. (2015). Selecting the number of trials in experimental biomechanics studies. *International Biomechanics*, *2*(1), 62–72. https://doi.org/10.1080/23335432.2015.1049296

Fox, E. L., & Mathews, D. K. (1974). *The Interval Training: Conditioning for Sports and General Fitness*. W.B. Saunders.

Friedman, J. H., Bentley, J. L., & Finkel, R. A. (1977). An Algorithm for Finding Best Matches in Logarithmic Expected Time. *ACM Transactions on Mathematical Software*, *3*(3), 209–226. https://doi.org/10.1145/355744.355745

Gathercole, R., Sporer, B., Stellingwerff, T., & Sleivert, G. (2015). Alternative Countermovement-Jump Analysis to Quantify Acute Neuromuscular Fatigue. *International Journal of Sports Physiology and Performance*, *10*(1), 84–92. https://doi.org/10.1123/ijspp.2013-0413

Geisser, S. (1975). The Predictive Sample Reuse Method with Applications. *Journal of the American Statistical Association*, *70*(350), 320–328. https://doi.org/10.1080/01621459.1975.10479865

Giroux, C., Rabita, G., Chollet, D., & Guilhem, G. (2014). What is the Best Method for Assessing Lower Limb Force-Velocity Relationship? *International Journal of Sports Medicine*, *36*(02), 143–149. https://doi.org/10.1055/s-0034-1385886

Glazier, P. S., Wheat, J. S., Pease, D. L., & Bartlett, R. M. (2006). The Interface of Biomechanics and Motor Control. In K. Davids, S. J. Bennett, & K. M. Newell (Eds.), *Movement System Variability* (pp. 49–69). Human Kinetics.

Godwin, A., Agnew, M., & Stevenson, J. (2009). Accuracy of Inertial Motion Sensors in Static, Quasistatic, and Complex Dynamic Motion. *Journal of Biomechanical Engineering*, *131*(11), 114501. https://doi.org/10.1115/1.4000109

González-Ravé, J. M., Arija, A., & Clemente-Suarez, V. (2011). Seasonal Changes in Jump Performance and Body Composition in Women Volleyball Players: *Journal of Strength and Conditioning Research*, *25*(6), 1492–1501. https://doi.org/10.1519/JSC.0b013e3181da77f6

Goodfellow, I. (2017). NIPS 2016 Tutorial: Generative Adversarial Networks. *ArXiv:1701.00160 [Cs]*. http://arxiv.org/abs/1701.00160

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Networks. *ArXiv:1406.2661 [Cs, Stat]*. http://arxiv.org/abs/1406.2661

Gower, J. C., & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, *3*(1), 5–48. https://doi.org/10.1007/BF01896809

Gramacy, R. B., Gray, G. A., Le Digabel, S., Lee, H. K. H., Ranjan, P., Wells, G., & Wild, S. M. (2016). Modeling an Augmented Lagrangian for Blackbox

Constrained Optimization. *Technometrics*, *58*(1), 1–11. https://doi.org/10.1080/00401706.2015.1014065

Gramacy, R. B., & Lee, H. K. H. (2011). Optimization Under Unknown Constraints*. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, & M. West (Eds.), *Bayesian Statistics 9* (pp. 229–256). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199694587.003.0008

Groh, B. H., Fleckenstein, M., & Eskofier, B. M. (2016). Wearable trick classification in freestyle snowboarding. *2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, 89–93. https://doi.org/10.1109/BSN.2016.7516238

Groh, B. H., Fleckenstein, M., Kautz, T., & Eskofier, B. M. (2017). Classification and visualization of skateboard tricks using wearable sensors. *Pervasive and Mobile Computing*, *40*, 42–55. https://doi.org/10.1016/j.pmcj.2017.05.007

Guo, Y., Storm, F., Zhao, Y., Billings, S., Pavic, A., Mazzà, C., & Guo, L.-Z. (2017). A New Proxy Measurement Algorithm with Application to the Estimation of Vertical Ground Reaction Forces Using Wearable Sensors. *Sensors*, *17*(10), 2181. https://doi.org/10.3390/s17102181

Gurchiek, R. D., Rupasinghe Arachchige Don, H. S., Pelawa Watagoda, L. C. R., McGinnis, R. S., van Werkhoven, H., Needle, A. R., McBride, J. M., & Arnholt, A. T. (2019). Sprint Assessment Using Machine Learning and a Wearable Accelerometer. *Journal of Applied Biomechanics*, *35*(2), 164–169. https://doi.org/10.1123/jab.2018-0107

Haibo He, Yang Bai, Garcia, E. A., & Shutao Li. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

Halilaj, E., Rajagopal, A., Fiterau, M., Hicks, J. L., Hastie, T. J., & Delp, S. L. (2018). Machine learning in human movement biomechanics: Best practices, common pitfalls, and new opportunities. *Journal of Biomechanics*, *81*, 1–11. https://doi.org/10.1016/j.jbiomech.2018.09.009

Hall, P., & Robinson, A. P. (2009). Reducing variability of crossvalidation for smoothing-parameter choice. *Biometrika*, *96*(1), 175–186. https://doi.org/10.1093/biomet/asn068

Hammerla, N. Y., Halloran, S., & Ploetz, T. (2016). Deep, Convolutional, and Recurrent Models for Human Activity Recognition using Wearables. *ArXiv:1604.08880 [Cs, Stat]*. http://arxiv.org/abs/1604.08880

Harman, E. A., Rosenstein, M. T., Frykman, P. N., & Rosenstein, R. M. (1990). The effects of arms and countermovement on vertical jumping: *Medicine & Science in Sports & Exercise*, *22*(6), 825. https://doi.org/10.1249/00005768-199012000-00015

Harman, E. A., Rosenstein, M. T., Frykman, P. N., Rosenstein, R. M., & Kraemer, W. (1991). Estimation of Human Power Output from Vertical Jump. *Journal of Strength and Conditioning Research*, *5*(3), 116–120.

Harrison, A. J. (2014). *Applications of Functional Data Analysis in Sport Biomechanics*. 9.

Harrison, A. J., Ryan, W., & Hayes, K. (2007). Functional data analysis of joint coordination in the development of vertical jump performance. *Sports Biomechanics*, *6*(2), 199–214. https://doi.org/10.1080/14763140701323042

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed). Springer.

Hatze, H. (1998). Validity and Reliability of Methods for Testing Vertical Jumping Performance. *Journal of Applied Biomechanics*, *14*(2), 127–140. https://doi.org/10.1123/jab.14.2.127

Helwig, N. E., Hong, S., Hsiao-Wecksler, E. T., & Polk, J. D. (2011). Methods to temporally align gait cycle data. *Journal of Biomechanics*, *44*(3), 561–566. https://doi.org/10.1016/j.jbiomech.2010.09.015

Hinton, G. E. (1990). Connectionist learning procedures. In *Machine Learning: Vol. III* (pp. 555–610). Elsevier. https://doi.org/10.1016/B978-0-08-051055-2.50029-8

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, *18*(7), 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527

Hochmuth, G., & Marhold, G. (1977). The further development of biomechanical principles. *Proceedings of the 6th International Conference of Biomechanics, Denmark*, *2B*, 93–106.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Hojka, V., Tufano, J. J., Maly, T., Stastny, P., Jebavy, R., Feher, J., Zahalka, F., & Gryc, T. (2018). Concurrent validity of Myotest for assessing explosive strength indicators in countermovement jump. *Acta Gymnica*, *48*(3), 95–102. https://doi.org/10.5507/ag.2018.013

Hopkins, W. (2016). *SAS (and R) for Mixed Models (workshop materials)*. sportsci.org. http://sportscience.sportsci.org/2016/Mixed-model%20Workshop.zip

Hopkins, W. G. (2000). Measures of Reliability in Sports Medicine and Science: *Sports Medicine*, *30*(1), 1–15. https://doi.org/10.2165/00007256-200030010-00001

Hopkins, W. G. (2004). How to Interpret Changes in an Athletic Performance Test. *Sportscience*, *8*, 1–7. sportsci.org/jour/04/wghtests.htm

Hopkins, W. G. (2015). *Spreadsheets for analysis of validity and reliability*. *19*, 36–42. http://www.sportsci.org/2015/ValidRely.htm

Hori, N., Newton, R. U., Andrews, W. A., Kawamori, N., Mcguigan, M. R., & Nosaka, K. (2007). Comparison of four different methods to measure power

during hang power clean and weighted jump squat. *Journal of Strength & Conditioning Research*, *21*(2), 314–320.

Isaksson, A., Wallman, M., Göransson, H., & Gustafsson, M. G. (2008). Cross-validation and bootstrapping are unreliable in small sample classification. *Pattern Recognition Letters*, *29*(14), 1960–1965. https://doi.org/10.1016/j.patrec.2008.06.018

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2018). Image-to-Image Translation with Conditional Adversarial Networks. *ArXiv:1611.07004 [Cs]*. http://arxiv.org/abs/1611.07004

Jacobs, R., Bobbert, M. F., & van Ingen Schenau, G. J. (1996). Mechanical output from individual muscles during explosive leg extensions: The role of biarticular muscles. *Journal of Biomechanics*, *29*(4), 513–523. https://doi.org/10.1016/0021-9290(95)00067-4

Japkowicz, N. (2000). The Class Imbalance Problem: Significance and Strategies. *Proceedings of the 2000 International Conference on Artificial Intelligence*, 7.

Jensen, J., Thelen, E., & Ulrich, B. D. (1989). Constraints on multi-joint movements: From the spontaneity of infancy to the skill of adults. *Human Movement Science*, *8*(4), 393–402. https://doi.org/10.1016/0167-9457(89)90044-4

Jensen, R., Ebben, W. P., Petushek, E. J., Moran, K., O'Connor, N. E., & Richter, C. (2013). Continuous waveform analysis of force, velocity, and power adaptations to a periodized plyometric training program. *Proceedings of XXXI Congress of the International Society of Biomechanics in Sports*.

Jiménez-Reyes, P., Samozino, P., Pareja-Blanco, F., Conceição, F., Cuadrado-Peñafiel, V., González-Badillo, J. J., & Morin, J.-B. (2017). Validity of a Simple Method for Measuring Force-Velocity-Power Profile in Countermovement Jump. *International Journal of Sports Physiology and Performance*, *12*(1), 36–43. https://doi.org/10.1123/IJSPP.2015-0484

Johnson, D., & Bahamonde, R. (1996). Power Output Estimate in University Athletes. *Journal of Strength and Conditioning Research*, *10*(3), 161–166.

Johnson, W., Mian, A., Donnelly, C., Lloyd, D., & Alderson, J. (2017). Prediction of ground reaction forces and moments via supervised learning is independent of participant sex, height and mass. *ISBS Proceedings Archive*, *35*. https://commons.nmu.edu/isbs/vol35/iss1/25

Johnson, W., Mian, A., Robinson, M. A., Verheul, J., Lloyd, D. G., & Alderson, J. A. (2019). Multidimensional ground reaction forces and moments from wearable sensor accelerations via deep learning. *ArXiv:1903.07221v2*, 18.

Johnston, R. D., Gibson, N. V., Twist, C., Gabbett, T. J., MacNay, S. A., & MacFarlane, N. G. (2013). Physiological Responses to an Intensified Period of Rugby League Competition: *Journal of Strength and Conditioning Research*, *27*(3), 643–654. https://doi.org/10.1519/JSC.0b013e31825bb469

Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed). Springer.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A:*

Mathematical, Physical and Engineering Sciences, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Jones, D., Schonlau, M., & Welch, W. J. (1998). Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization*, *13*(4), 455–492. https://doi.org/10.1023/A:1008306431147

Jones, T., Smith, A., Macnaughton, L. S., & French, D. N. (2016). Strength and Conditioning and Concurrent Training Practices in Elite Rugby Union: *Journal of Strength and Conditioning Research*, *30*(12), 3354–3366. https://doi.org/10.1519/JSC.0000000000001445

Kautz, T., Groh, B. H., Hannink, J., Jensen, U., Strubberg, H., & Eskofier, B. M. (2017). Activity recognition in beach volleyball using a Deep Convolutional Neural Network: Leveraging the potential of Deep Learning in sports. *Data Mining and Knowledge Discovery*, *31*(6), 1678–1705. https://doi.org/10.1007/s10618-017-0495-0

Kelso, J. A. S. (1995). *Dynamic patterns: The self-organization of brain and behavior.* (pp. xvii, 334). The MIT Press.

Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. *Proceedings of ICNN'95 - International Conference on Neural Networks*, *4*, 1942–1948. https://doi.org/10.1109/ICNN.1995.488968

Kennedy, R., & Drake, D. (2018). Is a Bimodal Force-Time Curve Related to Countermovement Jump Performance? *Sports*, *6*(2), 36. https://doi.org/10.3390/sports6020036

Kibele, A. (1998). Possibilities and Limitations in the Biomechanical Analysis of Countermovement Jumps: A Methodological Study. *Journal of Applied Biomechanics*, *14*(1), 105–117. https://doi.org/10.1123/jab.14.1.105

Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, *53*(11), 3735–3745. https://doi.org/10.1016/j.csda.2009.04.009

Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. *ArXiv:1312.6114 [Cs, Stat]*. http://arxiv.org/abs/1312.6114

Kipp, K., & Harris, C. (2015). Patterns of barbell acceleration during the snatch in weightlifting competition. *Journal of Sports Sciences*, *33*(14), 1467–1471. https://doi.org/10.1080/02640414.2014.992035

Kipp, K., Redden, J., Sabick, M. B., & Harris, C. (2012a). Weightlifting Performance Is Related to Kinematic and Kinetic Patterns of the Hip and Knee Joints: *Journal of Strength and Conditioning Research*, *26*(7), 1838–1844. https://doi.org/10.1519/JSC.0b013e318239c1d2

Kipp, K., Redden, J., Sabick, M., & Harris, C. (2012b). Kinematic and Kinetic Synergies of the Lower Extremities During the Pull in Olympic Weightlifting. *Journal of Applied Biomechanics*, *28*(3), 271–278. https://doi.org/10.1123/jab.28.3.271

Kirby, T. J., McBride, J. M., Haines, T. L., & Dayne, A. M. (2011). Relative Net Vertical Impulse Determines Jumping Performance. *Journal of Applied Biomechanics*, *27*(3), 207–214. https://doi.org/10.1123/jab.27.3.207

Kneip, A., & Ramsay, J. O. (2008). Combining Registration and Fitting for Functional Models. *Journal of the American Statistical Association*, *103*(483), 1155–1165. https://doi.org/10.1198/016214508000000517

Knudson, D. V. (2009). Correcting the Use of the Term "Power" in the Strength and Conditioning Literature: *Journal of Strength and Conditioning Research*, *23*(6), 1902–1908. https://doi.org/10.1519/JSC.0b013e3181b7f5e5

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1137–1143.

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, *30*.

Krstajic, D., Buturovic, L. J., Leahy, D. E., & Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics*, *6*(1), 10. https://doi.org/10.1186/1758-2946-6-10

Kubat, M., & Matwin, S. (1997). Addressing the Curse of Imbalanced Training Sets: One Sided Selection. *Proceedings of the Fourteenth International Conference on Machine Learning*, 179–186.

Kuzmits, F. E., & Adams, A. J. (2008). The NFL Combine: Does It Predict Performance in the National Football League?: *Journal of Strength and Conditioning Research*, *22*(6), 1721–1727. https://doi.org/10.1519/JSC.0b013e318185f09d

Lafortune, M. A., Henning, E., & Valiant, G. A. (1995). Tibial shock measured with bone and skin mounted transducers. *Journal of Biomechanics*, *28*(8), 989–993. https://doi.org/10.1016/0021-9290(94)00150-3

Lake, J., & McMahon, J. (2018). Within-Subject Consistency of Unimodal and Bimodal Force Application during the Countermovement Jump. *Sports*, *6*(4), 143. https://doi.org/10.3390/sports6040143

Lara, A. J., Abián, J., Alegre, L. M., Jiménez, L., & Aguado, X. (2006). Assessment of power output in jump tests for applicants to a sports sciences degree. *The Journal of Sports Medicine and Physical Fitness*, *46*(3), 419–424.

Lara-Sánchez, A. J., Zagalaz, M. L., Berdejo-del-Fresno, D., & Martínez-López, E. J. (2011). Jump Peak Power Assessment Through Power Prediction Equations in Different Samples: *Journal of Strength and Conditioning Research*, *25*(7), 1957–1962. https://doi.org/10.1519/JSC.0b013e3181e06ef8

Latash, M. L. (2012). The bliss (not the problem) of motor abundance (not redundancy). *Experimental Brain Research*, *217*(1), 1–5. https://doi.org/10.1007/s00221-012-3000-4

Leard, J. S., Cirillo, M. A., Katsnelson, E., Kimiatek, D. A., Miller, T. W., Trebincevic, K., & Garbalosa, J. C. (2007). Validity of two alternative systems for measuring vertical jump height. *Journal of Strength and Conditioning Research*, *21*(4), 1296–1299. https://doi.org/10.1519/R-21536.1

LeCun, Y., Bottou, L., Orr, G. B., & Müller, K.-R. (1998). Efficient BackProp. In G. B. Orr & K.-R. Müller (Eds.), *Neural Networks: Tricks of the Trade* (Vol.

1524, pp. 9–50). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-49430-8_2

Lesinski, M., Muehlbauer, T., & Granacher, U. (2016). Concurrent validity of the Gyko inertial sensor system for the assessment of vertical jump height in female sub-elite youth soccer players. *BMC Sports Science, Medicine and Rehabilitation*, *8*(1), 35. https://doi.org/10.1186/s13102-016-0061-x

Letham, B., Karrer, B., Ottoni, G., & Bakshy, E. (2019). Constrained Bayesian Optimization with Noisy Experiments. *Bayesian Analysis*, *14*(2), 495–519. https://doi.org/10.1214/18-BA1110

Levitin, D. J., Nuzzo, R. L., Vines, B. W., & Ramsay, J. O. (2007). Introduction to functional data analysis. *Canadian Psychology/Psychologie Canadienne*, *48*(3), 135–155. https://doi.org/10.1037/cp2007014

Li, C., Gupta, S., Rana, S., Nguyen, V., Venkatesh, S., & Shilton, A. (2017). High Dimensional Bayesian Optimization using Dropout. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2096–2102. https://doi.org/10.24963/ijcai.2017/291

Liebl, D., Willwacher, S., Hamill, J., & Brüggemann, G.-P. (2014). Ankle plantarflexion strength in rearfoot and forefoot runners: A novel clusteranalytic approach. *Human Movement Science*, *35*, 104–120. https://doi.org/10.1016/j.humov.2014.03.008

Lin, Z., Wang, L., & Cao, J. (2016). Interpretable functional principal component analysis: Interpretable Functional Principal Component Analysis. *Biometrics*, *72*(3), 846–854. https://doi.org/10.1111/biom.12457

Lombardi, M. (2008). The Accuracy & Stability of Quartz Watches. *Horological Journal*, *February*, 57–59.

Loturco, I., Pereira, L. A., Cal Abad, C. C., D'Angelo, R. A., Fernandes, V., Kitamura, K., Kobal, R., & Nakamura, F. Y. (2015). Vertical and Horizontal Jump Tests Are Strongly Associated With Competitive Performance in 100-m Dash Events: *Journal of Strength and Conditioning Research*, *29*(7), 1966–1971. https://doi.org/10.1519/JSC.0000000000000849

Luhtanen, P., & Komi, P. V. (1978). Segmental contribution to forces in vertical jump. *European Journal of Applied Physiology and Occupational Physiology*, *38*(3), 181–188. https://doi.org/10.1007/BF00430076

Luinge, H. J., & Veltink, P. H. (2005). Measuring orientation of human body segments using miniature gyroscopes and accelerometers. *Medical & Biological Engineering & Computing*, *43*(2), 273–282. https://doi.org/10.1007/BF02345966

Luo, G. (2016). A review of automatic selection methods for machine learning algorithms and hyper-parameter values. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *5*(1), 18. https://doi.org/10.1007/s13721-016-0125-6

Madgwick, S. O. H., Harrison, A. J. L., & Vaidyanathan, R. (2011). Estimation of IMU and MARG orientation using a gradient descent algorithm. *2011 IEEE International Conference on Rehabilitation Robotics*, 1–7. https://doi.org/10.1109/ICORR.2011.5975346

Madigan, D., & Raftery, A. E. (1994). Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. *Journal of the American Statistical Association*, *89*(428), 1535–1546. https://doi.org/10.1080/01621459.1994.10476894

Maiwald, C., Dannemann, A., Gaudel, J., & Oriwol, D. (2015). A simple method to detect stride intervals in continuous acceleration and gyroscope data recorded during treadmill running. *Footwear Science*, *7*(sup1), S143–S144. https://doi.org/10.1080/19424280.2015.1038656

Makaruk, H., Winchester, J. B., Sadowski, J., Czaplicki, A., & Sacewicz, T. (2011). Effects of Unilateral and Bilateral Plyometric Training on Power and Jumping Ability in Women: *Journal of Strength and Conditioning Research*, *25*(12), 3311–3318. https://doi.org/10.1519/JSC.0b013e318215fa33

Mallat, S. (2012). Group Invariant Scattering. *ArXiv:1101.2286 [Cs, Math]*. http://arxiv.org/abs/1101.2286

Mallor, F., Leon, T., Gaston, M., & Izquierdo, M. (2010). Changes in power curve shapes as an indicator of fatigue during dynamic contractions. *Journal of Biomechanics*, *43*(8), 1627–1631. https://doi.org/10.1016/j.jbiomech.2010.01.038

Mandic, R., Jakovljevic, S., & Jaric, S. (2015). Effects of countermovement depth on kinematic and kinetic patterns of maximum vertical jumps. *Journal of Electromyography and Kinesiology*, *25*(2), 265–272. https://doi.org/10.1016/j.jelekin.2014.11.001

Mandic, R., Knezevic, O. M., Mirkov, D. M., & Jaric, S. (2016). Control strategy of maximum vertical jumps: The preferred countermovement depth may not be fully optimized for jump height. *Journal of Human Kinetics*, *52*(1), 85–94. https://doi.org/10.1515/hukin-2015-0196

Marcora, S., & Miller, M. K. (2000). The effect of knee angle on the external validity of isometric measures of lower body neuromuscular function. *Journal of Sports Sciences*, *18*(5), 313–319. https://doi.org/10.1080/026404100402377

Markovic, G., Dizdar, D., Jukic, I., & Cardinale, M. (2004). Reliability and factorial validity of squat and countermovement jump tests. *Journal of Strength & Conditioning Research*, *18*(3), 551–555.

Markovic, G., Vuk, S., & Jaric, S. (2011). Effects of Jump Training with Negative versus Positive Loading on Jumping Mechanics. *International Journal of Sports Medicine*, *32*(05), 365–372. https://doi.org/10.1055/s-0031-1271678

Markovic, S., Mirkov, D. M., Knezevic, O. M., & Jaric, S. (2013). Jump training with different loads: Effects on jumping performance and power output. *European Journal of Applied Physiology*, *113*(10), 2511–2521. https://doi.org/10.1007/s00421-013-2688-6

Marouani, H., & Dagenais, M. R. (2008). *Internal Clock Drift Estimation in Computer Clusters*. *2008*. https://doi.org/10.1155/2008/583162

Marquardt, D. W., & Snee, R. D. (1975). Ridge Regression in Practice. *The American Statistician*, *29*(1), 3–20. https://doi.org/10.1080/00031305.1975.10479105

Marrier, B., Le Meur, Y., Robineau, J., Lacome, M., Couderc, A., Hausswirth, C., Piscione, J., & Morin, J.-B. (2017). Quantifying Neuromuscular Fatigue Induced by an Intense Training Session in Rugby Sevens. *International Journal of Sports Physiology and Performance*, *12*(2), 218–223. https://doi.org/10.1123/ijspp.2016-0030

Marron, J. S., Ramsay, J. O., Sangalli, L. M., & Srivastava, A. (2015). Functional Data Analysis of Amplitude and Phase Variation. *Statistical Science*, *30*(4), 468–484. https://doi.org/10.1214/15-STS524

Marshall, B., Franklyn-Miller, A., Moran, K., King, E., Richter, C., Gore, S., Strike, S., & Falvey, É. (2015). Biomechanical symmetry in elite rugby union players during dynamic tasks: An investigation using discrete and continuous data analysis techniques. *BMC Sports Science, Medicine and Rehabilitation*, *7*(1), 13. https://doi.org/10.1186/s13102-015-0006-9

Mauch, M., Rist, H.-J., & Kaelin, X. (2014). Reliability and Validity of Two Measurement Systems in the Quantification of Jump Performance. *Swiss Sports & Exercise Medicine*, *62*(1). https://doi.org/10.34045/SSEM/2014/6

Mazzà, C., Donati, M., McCamley, J., Picerno, P., & Cappozzo, A. (2012). An optimized Kalman filter for the estimate of trunk orientation from inertial sensors data during treadmill walking. *Gait & Posture*, *35*(1), 138–142. https://doi.org/10.1016/j.gaitpost.2011.08.024

McErlain-Naylor, S., King, M., & Pain, M. T. G. (2014). Determinants of countermovement jump performance: A kinetic and kinematic analysis. *Journal of Sports Sciences*, *32*(19), 1805–1812. https://doi.org/10.1080/02640414.2014.924055

McGee, K. J., & Burkett, L. N. (2003). The National Football League combine: A reliable predictor of draft status? *Journal of Strength and Conditioning Research*, *17*(1), 6–11.

McGrath, J. W., Neville, J., Stewart, T., & Cronin, J. (2019). Cricket fast bowling detection in a training setting using an inertial measurement unit and machine learning. *Journal of Sports Sciences*, *37*(11), 1220–1226. https://doi.org/10.1080/02640414.2018.1553270

McLean, B. D., Coutts, A. J., Kelly, V., McGuigan, M. R., & Cormack, S. J. (2010). Neuromuscular, Endocrine, and Perceptual Fatigue Responses During Different Length Between-Match Microcycles in Professional Rugby League Players. *International Journal of Sports Physiology and Performance*, *5*(3), 367–383. https://doi.org/10.1123/ijspp.5.3.367

McLellan, C. P., Lovell, D. I., & Gass, G. C. (2011). The Role of Rate of Force Development on Vertical Jump Performance: *Journal of Strength and Conditioning Research*, *25*(2), 379–385. https://doi.org/10.1519/JSC.0b013e3181be305c

McMahon, J. J., Jones, P. A., Suchomel, T. J., Lake, J., & Comfort, P. (2018). Influence of the Reactive Strength Index Modified on Force– and Power–Time Curves. *International Journal of Sports Physiology and Performance*, *13*(2), 220–227. https://doi.org/10.1123/ijspp.2017-0056

McMahon, J., Lake, J. P., & Suchomel, T. J. (2019). Vertical jump testing. In *Performance Assessment in Strength and Conditioning*. Routledge, Taylor & Francis Group.

McMahon, J., Murphy, S., Rej, S. J. E., & Comfort, P. (2017). Countermovement-Jump-Phase Characteristics of Senior and Academy Rugby League Players. *International Journal of Sports Physiology and Performance*, *12*(6), 803–811. https://doi.org/10.1123/ijspp.2016-0467

Mercer, J., Bates, B., Dufek, J., & Hreljac, A. (2003). Characteristics of shock attenuation during fatigued running. *Journal of Sports Sciences*, *21*(11), 911–919. https://doi.org/10.1080/0264041031000140383

Meyer, U., Ernst, D., Schott, S., Riera, C., Hattendorf, J., Romkes, J., Granacher, U., Göpfert, B., & Kriemler, S. (2015). Validation of two accelerometers to determine mechanical loading of physical activities in children. *Journal of Sports Sciences*, *33*(16), 1702–1709. https://doi.org/10.1080/02640414.2015.1004638

Meylan, C. M. P., Nosaka, K., Green, J., & Cronin, J. B. (2011). The Effect of Three Different Start Thresholds on the Kinematics and Kinetics of a Countermovement Jump: *Journal of Strength and Conditioning Research*, *25*(4), 1164–1167. https://doi.org/10.1519/JSC.0b013e3181c699b9

Miller, D. I., & East, D. J. (1976). Kinematic and kinetic correlates of vertical jumping in woman. In P. V. Komi (Ed.), *Biomechanics V-B* (pp. 65–72). Baltimore, University Park Press.

Mlakar, M., & Luštrek, M. (2017). Analyzing tennis game through sensor data with machine learning and multi-objective optimization. *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, 153–156. https://doi.org/10.1145/3123024.3123163

Mockus, J. (1977). On Bayesian Methods for Seeking the Extremum and their Application. *IFIP Congress*.

Moir, G., Button, C., Glaister, M., & Stone, M. H. (2004). *Influence of familiarization on the reliability of vertical jump and acceleration sprinting performance in physically active men*. *18*(2), 276–280.

Molinaro, A. M., Simon, R., & Pfeiffer, R. M. (2005). Prediction error estimation: A comparison of resampling methods. *Bioinformatics*, *21*(15), 3301–3307. https://doi.org/10.1093/bioinformatics/bti499

Monnet, T., Decatoire, A., & Lacouture, P. (2014). Comparison of algorithms to determine jump height and flight time from body mounted accelerometers. *Sports Engineering*, *17*(4), 249–259. https://doi.org/10.1007/s12283-014-0155-1

Moritz, H. (2000). Geodetic Reference System 1980. *Journal of Geodesy*, *74*(1), 128–133. https://doi.org/10.1007/s001900050278

Mosteller, F., & Tukey, J. W. (1968). Data analysis, including statistics. *Handbook of Social Psychology*, *2*, 80–203.

Moudy, S., Richter, C., & Strike, S. (2018). Landmark registering waveform data improves the ability to predict performance measures. *Journal of Biomechanics*, *78*, 109–117. https://doi.org/10.1016/j.jbiomech.2018.07.027

Mujika, I., Santisteban, J., Impellizzeri, F. M., & Castagna, C. (2009). Fitness determinants of success in men's and women's football. *Journal of Sports Sciences*, *27*(2), 107–114. https://doi.org/10.1080/02640410802428071

Mundt, M., Koeppe, A., David, S., Witter, T., Bamer, F., Potthast, W., & Markert, B. (2020). Estimation of Gait Mechanics Based on Simulated and Measured IMU Data Using an Artificial Neural Network. *Frontiers in Bioengineering and Biotechnology*, *8*, 41. https://doi.org/10.3389/fbioe.2020.00041

Myers, D. C., Gebhardt, D. L., Crump, C. E., & Fleishman, E. A. (1993). The Dimensions of Human Physical Performance: Factor Analysis of Strength, Stamina, Flexibility, and Body Composition Measures. *Human Performance*, *6*(4), 309–344. https://doi.org/10.1207/s15327043hup0604_2

Nagano, A., Komura, T., & Fukashiro, S. (2007). Optimal coordination of maximal-effort horizontal and vertical jump motions – a computer simulation study. *BioMedical Engineering OnLine*, *6*(1), 20. https://doi.org/10.1186/1475-925X-6-20

Neugebauer, J. M., Collins, K. H., & Hawkins, D. A. (2014). Ground Reaction Force Estimates from ActiGraph GT3X+ Hip Accelerations. *PLoS ONE*, *9*(6), e99023. https://doi.org/10.1371/journal.pone.0099023

Neugebauer, J. M., Hawkins, D. A., & Beckett, L. (2012). Estimating Youth Locomotion Ground Reaction Forces Using an Accelerometer-Based Activity Monitor. *PLoS ONE*, *7*(10), e48182. https://doi.org/10.1371/journal.pone.0048182

Newton, R. U., Rogers, R. A., Volek, J. S., Häkkinen, K., & Kraemer, W. J. (2006). Four Weeks of Optimal Load Ballistic Resistance Training at the End of Season Attenuates Declining Jump Performance of Women Volleyball Players. *The Journal of Strength and Conditioning Research*, *20*(4), 955. https://doi.org/10.1519/R-5050502x.1

Ngoh, K. J.-H., Gouwanda, D., Gopalai, A. A., & Chong, Y. Z. (2018). Estimation of vertical ground reaction force during running using neural network model and uniaxial accelerometer. *Journal of Biomechanics*, *76*, 269–273. https://doi.org/10.1016/j.jbiomech.2018.06.006

Nilsson, J., & Thorstensson, A. (1989). Ground reaction forces at different speeds of human walking and running. *Acta Physiologica Scandinavica*, *136*(2), 217–227. https://doi.org/10.1111/j.1748-1716.1989.tb08655.x

Oddsson, L. (1987). What Factors Determine Vertical Jumping Height? *5th International Symposium in Biomechanics in Sports*, *5*, 393–401.

Olmschenk, G., Zhu, Z., & Tang, H. (2019). Generalizing semi-supervised generative adversarial networks to regression using feature contrasting. *Computer Vision and Image Understanding*, *186*, 1–12. https://doi.org/10.1016/j.cviu.2019.06.004

Olson, R. S., Bartley, N., Urbanowicz, R. J., & Moore, J. H. (2016). Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science.

Proceedings of the Genetic and Evolutionary Computation Conference 2016, 485–492. https://doi.org/10.1145/2908812.2908918

Ordóñez, F., & Roggen, D. (2016). Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors*, *16*(1), 115. https://doi.org/10.3390/s16010115

O'Reilly, M. A., Whelan, D. F., Ward, T. E., Delahunt, E., & Caulfield, B. (2017). Classification of lunge biomechanics with multiple and individual inertial measurement units. *Sports Biomechanics*, *16*(3), 342–360. https://doi.org/10.1080/14763141.2017.1314544

Owen, N. J., Watkins, J., Kilduff, L. P., Bevan, H. R., & Bennett, M. A. (2014). Development of a Criterion Method to Determine Peak Mechanical Power Output in a Countermovement Jump: *Journal of Strength and Conditioning Research*, *28*(6), 1552–1558. https://doi.org/10.1519/JSC.0000000000000311

Page, A., & Epifanio, I. (2007). A simple model to analyze the effectiveness of linear time normalization to reduce variability in human movement analysis. *Gait & Posture*, *25*(1), 153–156. https://doi.org/10.1016/j.gaitpost.2006.01.006

Pan, S. J., & Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Payne, A. H., Slater, W. J., & Telford, T. (1968). The Use of a Force Platform in the Study of Athletic Activities. A Preliminary Investigation. *Ergonomics*, *11*(2), 123–143. https://doi.org/10.1080/00140136808930950

Perrine, J. J., Gregor, R., Munroe, R., & Edgerton, V. R. (1978). Muscle power capacities and temporal output patterns of skilled vs. Non-skilled vertical jumpers. *Medicine & Science in Sports & Exercise*, *10*, 64.

Picard, R. R., & Cook, R. D. (1984). Cross-Validation of Regression Models. *Journal of the American Statistical Association*, *79*(387), 575–583. https://doi.org/10.1080/01621459.1984.10478083

Picerno, P., Camomilla, V., & Capranica, L. (2011). Countermovement jump performance assessment using a wearable 3D inertial measurement unit. *Journal of Sports Sciences*, *29*(2), 139–146. https://doi.org/10.1080/02640414.2010.523089

Picerno, P., Cereatti, A., & Cappozzo, A. (2011). A spot check for assessing static orientation consistency of inertial and magnetic sensing units. *Gait & Posture*, *33*(3), 373–378. https://doi.org/10.1016/j.gaitpost.2010.12.006

Picheny, V., Ginsbourger, D., Richet, Y., & Caplin, G. (2013). Quantile-Based Optimization of Noisy Computer Experiments With Tunable Precision. *Technometrics*, *55*(1), 2–13. https://doi.org/10.1080/00401706.2012.707580

Picheny, V., Wagner, T., & Ginsbourger, D. (2013). A benchmark of kriging-based infill criteria for noisy optimization. *Structural and Multidisciplinary Optimization*, *48*(3), 607–626. https://doi.org/10.1007/s00158-013-0919-4

Pion, J. A., Fransen, J., Deprez, D. N., Segers, V. I., Vaeyens, R., Philippaerts, R. M., & Lenoir, M. (2015). Stature and Jumping Height Are Required in Female Volleyball, but Motor Coordination Is a Key Factor for Future Elite Success:

Journal of Strength and Conditioning Research, *29*(6), 1480–1485. https://doi.org/10.1519/JSC.0000000000000778

Pogson, M., Verheul, J., Robinson, M. A., Vanrenterghem, J., & Lisboa, P. (2020). A neural network method to predict task- and step-specific ground reaction force magnitudes from trunk accelerations during running activities. *Medical Engineering & Physics*, *78*, 82–89. https://doi.org/10.1016/j.medengphy.2020.02.002

Preatoni, E., Hamill, J., Harrison, A. J., Hayes, K., Van Emmerik, R. E. A., Wilson, C., & Rodano, R. (2013). Movement variability and skills monitoring in sports. *Sports Biomechanics*, *12*(2), 69–92. https://doi.org/10.1080/14763141.2012.738700

Quagliarella, L., Sasanelli, N., Belgiovine, G., Moretti, L., & Moretti, B. (2010). Evaluation of Standing Vertical Jump by Ankles Acceleration Measurement: *Journal of Strength and Conditioning Research*, *24*(5), 1229–1236. https://doi.org/10.1519/JSC.0b013e3181cb281a

Quagliarella, L., Sasanelli, N., Belgiovine, G., Moretti, L., & Moretti, B. (2011). Power Output Estimation in Vertical Jump Performed by Young Male Soccer Players: *Journal of Strength and Conditioning Research*, *25*(6), 1638–1646. https://doi.org/10.1519/JSC.0b013e3181d85a99

Ramsay, J. O. (2017). *FDA Matlab Code Library*. http://www.psych.mcgill.ca/misc/fda/software.html

Ramsay, J. O., & Li, X. (1998). Curve registration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *60*(2), 351–363. https://doi.org/10.1111/1467-9868.00129

Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis* (2nd ed). Springer.

Rasmussen, C. E. (1999). *Evaluation of Gaussian processes and other methods for non-linear regression.* National Library of Canada = Bibliothèque nationale du Canada.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Rawashdeh, S. A., Rafeldt, D. A., Uhl, T. L., & Lumpp, J. E. (2015). Wearable motion capture unit for shoulder injury prevention. *2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, 1–6. https://doi.org/10.1109/BSN.2015.7299417

Requena, B., García, I., Requena, F., Saez-Saez de Villarreal, E., & Pääsuke, M. (2012). Reliability and validity of a wireless microelectromechanicals based system (keimove^{TM}) for measuring vertical jumping performance. *Journal of Sports Science & Medicine*, *11*(1), 115–122.

Rezagholiradeh, M., & Haidar, M. A. (2018). Reg-Gan: Semi-Supervised Learning Based on Generative Adversarial Networks for Regression. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2806–2810. https://doi.org/10.1109/ICASSP.2018.8462534

Richter, C., Gualano, L., O'Connor, N. E., & Moran, K. (2013). Cross-comparison of the performance of discrete, phase and functional data analysis to describe

dependent variable. *Proceedings of 31st International Conference on Biomechanics in Sports*.

Richter, C., O'Connor, N. E., Marshall, B., & Moran, K. (2014a). Analysis of Characterizing Phases on Waveforms: An Application to Vertical Jumps. *Journal of Applied Biomechanics*, *30*(2), 316–321. https://doi.org/10.1123/jab.2012-0218

Richter, C., O'Connor, N. E., Marshall, B., & Moran, K. (2014b). Comparison of discrete-point vs. Dimensionality-reduction techniques for describing performance-related aspects of maximal vertical jumping. *Journal of Biomechanics*, *47*(12), 3012–3017. https://doi.org/10.1016/j.jbiomech.2014.07.001

Robertson, D. G. E., & Fleming, D. (1987). Robertson (1987)—Kinetics of standing broad and vertical jumping.pdf. *Canadian Journal of Applied Sport Sciences.*, *12*(1), 19–23.

Rodríguez, J. D., Perez, A., & Lozano, J. A. (2010). Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(3), 569–575. https://doi.org/10.1109/TPAMI.2009.187

Rodríguez, J. D., Pérez, A., & Lozano, J. A. (2013). A general framework for the statistical analysis of the sources of variance for classification error estimators. *Pattern Recognition*, *46*(3), 855–864. https://doi.org/10.1016/j.patcog.2012.09.007

Roetenberg, D., Luinge, H., & Slycke, P. (2013). *Xsens MVN: Full 6DOF Human Motion Tracking Using Miniature Inertial Sensors*. 10.

Roetenberg, D., Slycke, P. J., & Veltink, P. H. (2007). Ambulatory Position and Orientation Tracking Fusing Magnetic and Inertial Sensing. *IEEE Transactions on Biomedical Engineering*, *54*(5), 883–890. https://doi.org/10.1109/TBME.2006.889184

Rowell, A. E., Aughey, R. J., Hopkins, W. G., Stewart, A. M., & Cormack, S. J. (2017). Identification of Sensitive Measures of Recovery After External Load From Football Match Play. *International Journal of Sports Physiology and Performance*, *12*(7), 969–976. https://doi.org/10.1123/ijspp.2016-0522

Ruddock, A. D., & Winter, E. M. (2016). Jumping depends on impulse not power. *Journal of Sports Sciences*, *34*(6), 584–585. https://doi.org/10.1080/02640414.2015.1064157

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Ryan, W., Harrison, A. J., & Hayes, K. (2006). Functional data analysis of knee joint kinematics in the vertical jump. *Sports Biomechanics*, *5*(1), 121–138. https://doi.org/10.1080/14763141.2006.9628228

Sabatini, A. M. (2006). Quaternion-Based Extended Kalman Filter for Determining Orientation by Inertial and Magnetic Sensing. *IEEE Transactions on*

Biomedical Engineering, *53*(7), 1346–1356. https://doi.org/10.1109/TBME.2006.875664

Sabatini, A. M. (2011). Estimating Three-Dimensional Orientation of Human Body Parts by Inertial/Magnetic Sensing. *Sensors*, *11*(2), 1489–1525. https://doi.org/10.3390/s110201489

Sabatini, A. M., Martelloni, C., Scapellato, S., & Cavallo, F. (2005). Assessment of Walking Features From Foot Inertial Sensing. *IEEE Transactions on Biomedical Engineering*, *52*(3), 486–494. https://doi.org/10.1109/TBME.2004.840727

Sacilotto, G. B. D., Warmenhoven, J. S., Mason, B. R., Ball, N., & Clothier, J. (2015). Investigation of ATM propulsion Force-time Profiles using Functional Data Analysis on Front Crawl Sprint Swimmers. *33rd International Conference of Biomechanics in Sports*.

Saha, S., & Lakes, R. S. (1977). The effect of soft tissue on wave-propagation and vibration tests for determining the in vivo properties of bone. In *Journal of Biomechanics* (Vol. 10, Issue 7, pp. 393–401). https://doi.org/10.1016/0021-9290(77)90015-X

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *26*(1), 43–49. https://doi.org/10.1109/TASSP.1978.1163055

Salles, A. S., Baltzopoulos, V., & Rittweger, J. (2011). Differential effects of countermovement magnitude and volitional effort on vertical jumping. *European Journal of Applied Physiology*, *111*(3), 441–448. https://doi.org/10.1007/s00421-010-1665-6

Samozino, P., Morin, J.-B., Hintzy, F., & Belli, A. (2008). A simple method for measuring force, velocity and power output during squat jump. *Journal of Biomechanics*, *41*(14), 2940–2945. https://doi.org/10.1016/j.jbiomech.2008.07.028

Sanders, R. T., & Wilson, B. D. (1992). Comparsion of static and counter movement jumps of unconstrained movement amplitude. *Australian Journal of Sports and Medicine in Sport*, *24*(3), 79–85.

Sapna, S. (2012). Backpropagation Learning Algorithm Based on Levenberg Marquardt Algorithm. *Computer Science & Information Technology (CS & IT)*, 393–398. https://doi.org/10.5121/csit.2012.2438

Sargent, D. A. (1921). The Physical Test of a Man. *American Physical Education Review*, *26*(4), 188–194. https://doi.org/10.1080/23267224.1921.10650486

Sayers, S. P., Harackiewicz, D. V., Harman, E. A., Frykman, P. N., & Rosenstein, M. T. (1999). Cross-validation of three jump power equations. *Medicine and Science in Sports and Exercise*, *31*(4), 572–577. https://doi.org/10.1097/00005768-199904000-00013

Senin, P. (2008). Dynamic time warping algorithm review. *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA*, *855*(1–23), 40.

Shao, J. (1992). Asymptotic theory in generalized ilinear models with nuisance scale parameters. *Probability Theory and Related Fields*, *91*(1), 25–41. https://doi.org/10.1007/BF01194488

Shao, J. (1993). Linear Model Selection by Cross-validation. *Journal of the American Statistical Association*, *88*(422), 486–494. https://doi.org/10.1080/01621459.1993.10476299

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, *7*(2), 221–242. https://www.jstor.org/stable/24306073

Sheppard, J. M., Cormack, S., Taylor, K.-L., McGuigan, M. R., & Newton, R. U. (2008). Assessing the Force-Velocity Characteristics of the Leg Extensors in Well-Trained Athletes: The Incremental Load Power Profile: *Journal of Strength and Conditioning Research*, *22*(4), 1320–1326. https://doi.org/10.1519/JSC.0b013e31816d671b

Shetty, A. B., & Etnyre, B. R. (1989). Contribution of Arm Movement to the Force Components of a Maximum Vertical Jump. *Journal of Orthopaedic & Sports Physical Therapy*, *11*(5), 198–201. https://doi.org/10.2519/jospt.1989.11.5.198

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, *6*(1), 60. https://doi.org/10.1186/s40537-019-0197-0

Shull, P. B., Jirattigalachote, W., Hunt, M. A., Cutkosky, M. R., & Delp, S. L. (2014). Quantified self and human movement: A review on the clinical impact of wearable sensing and feedback for gait analysis and intervention. *Gait & Posture*, *40*(1), 11–19. https://doi.org/10.1016/j.gaitpost.2014.03.189

Shull, P. B., Xu, J., Yu, B., & Zhu, X. (2017). Magneto-Gyro Wearable Sensor Algorithm for Trunk Sway Estimation During Walking and Running Gait. *IEEE Sensors Journal*, *17*(2), 480–486. https://doi.org/10.1109/JSEN.2016.2630938

Simon, R., Radmacher, M. D., Dobbin, K., & McShane, L. M. (2003). Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification. *JNCI Journal of the National Cancer Institute*, *95*(1), 14–18. https://doi.org/10.1093/jnci/95.1.14

Simons, C., & Bradshaw, E. J. (2016). Do accelerometers mounted on the back provide a good estimate of impact loads in jumping and landing tasks? *Sports Biomechanics*, *15*(1), 76–88. https://doi.org/10.1080/14763141.2015.1123765

Sinclair, J., Hobbs, S. J., Protheroe, L., Edmundson, C. J., & Greenhalgh, A. (2013). Determination of Gait Events Using an Externally Mounted Shank Accelerometer. *Journal of Applied Biomechanics*, *29*(1), 118–122. https://doi.org/10.1123/jab.29.1.118

Smirniotou, A., Katsikas, C., Paradisis, G., Argeitaki, P., Zacharogiannis, E., & Tziortzis, S. (2008). Strength-power parameters as predictors of sprinting performance. *The Journal of Sports Medicine and Physical Fitness*, *48*(4), 447–454.

Smith, K. E., & Smith, A. O. (2020). Conditional GAN for timeseries generation. *ArXiv:2006.16477 [Cs, Stat]*. http://arxiv.org/abs/2006.16477

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*(3), 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 25* (pp. 2951–2959). Curran Associates, Inc. http://papers.nips.cc/paper/4522-practical-bayesian-optimization-of-machine-learning-algorithms.pdf

Sole, C. J. (2015). *Analysis of Countermovement Vertical Jump Force-Time Curve Phase Characteristics in Athletes* [Electronic Theses and Dissertations, East Tennessee State University]. https://dc.etsu.edu/etd/2549

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*(2), 111–133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x

Street, G., McMillan, S., Board, W., Rasmussen, M., & Heneghan, J. M. (2001). Sources of Error in Determining Countermovement Jump Height with the Impulse Method. *Journal of Applied Biomechanics*, *17*(1), 43–54. https://doi.org/10.1123/jab.17.1.43

Taddy, M. A., Lee, H. K. H., Gray, G. A., & Griffin, J. D. (2009). Bayesian Guided Pattern Search for Robust Local Optimization. *Technometrics*, *51*(4), 389–401. https://doi.org/10.1198/TECH.2009.08007

Tan, T., Chiasson, D. P., Hu, H., & Shull, P. B. (2019). Influence of IMU position and orientation placement errors on ground reaction force estimation. *Journal of Biomechanics*, *97*, 109416. https://doi.org/10.1016/j.jbiomech.2019.109416

Tanaka, F. H. K. dos S., & Aranha, C. (2019). Data Augmentation Using GANs. *ArXiv:1904.09135 [Cs, Stat]*. http://arxiv.org/abs/1904.09135

Taylor, K.-L., Chapman, D. W., Cronin, J. B., Newton, M. J., & Gill, N. (2012). Fatigue monitoring in High Performance Sport: A Survey of Current Trends. *Journal of Australian Strength & Conditioning*, *20*(1), 12.

Taylor, K.-L., Cronin, J., Gill, N. D., Chapman, D. W., & Sheppard, J. (2010). Sources of Variability in Iso-Inertial Jump Assessments. *International Journal of Sports Physiology and Performance*, *5*(4), 546–558. https://doi.org/10.1123/ijspp.5.4.546

Tessier, J.-F., Basset, F.-A., Simoneau, M., & Teasdale, N. (2013). Lower-Limb Power cannot be Estimated Accurately from Vertical Jump Tests. *Journal of Human Kinetics*, *38*, 5–13. https://doi.org/10.2478/hukin-2013-0040

Tian, Y., Wei, H., & Tan, J. (2013). An Adaptive-Gain Complementary Filter for Real-Time Human Motion Tracking With MARG Sensors in Free-Living Environments. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *21*(2), 254–264. https://doi.org/10.1109/TNSRE.2012.2205706

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tibshirani, R. (2020). *Supervised Principal Components for regression and survival analysis.* (1.12) [Computer software]. https://github.com/jedazard/superpc

Tibshirani, R. J., & Tibshirani, R. (2009). A bias correction for the minimum error rate in cross-validation. *The Annals of Applied Statistics*, *3*(2), 822–829. https://doi.org/10.1214/08-AOAS224

Tillin, N. A., Pain, M. T. G., & Folland, J. (2013). Explosive force production during isometric squats correlates with athletic performance in rugby union players. *Journal of Sports Sciences*, *31*(1), 66–76. https://doi.org/10.1080/02640414.2012.720704

Torgo, L., Branco, P., Ribeiro, R. P., & Pfahringer, B. (2015). Resampling strategies for regression. *Expert Systems*, *32*(3), 465–476. https://doi.org/10.1111/exsy.12081

Torgo, L., & Ribeiro, R. (2007). Utility-Based Regression. In J. N. Kok, J. Koronacki, R. Lopez de Mantaras, S. Matwin, D. Mladenič, & A. Skowron (Eds.), *Knowledge Discovery in Databases: PKDD 2007* (Vol. 4702, pp. 597–604). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-74976-9_63

Torgo, L., Ribeiro, R. P., Pfahringer, B., & Branco, P. (2013). SMOTE for Regression. In L. Correia, L. P. Reis, & J. Cascalho (Eds.), *Progress in Artificial Intelligence* (Vol. 8154, pp. 378–389). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-40669-0_33

Tran, J., Netto, K., Aisbett, B., & Gastin, P. (2010). Validation of accelerometer data for measuring impacts during jumping and landing tasks. *Proceedings of the 28th International Conference on Biomechanics in Sports*, 1–4.

Twist, C., & Highton, J. (2013). Monitoring Fatigue and Recovery in Rugby League Players. *International Journal of Sports Physiology and Performance*, *8*(5), 467–474. https://doi.org/10.1123/ijspp.8.5.467

Ullah, S., & Finch, C. F. (2013). Applications of functional data analysis: A systematic review. *BMC Medical Research Methodology*, *13*(1), 43. https://doi.org/10.1186/1471-2288-13-43

Unler, A., & Murat, A. (2010). A discrete particle swarm optimization method for feature selection in binary classification problems. *European Journal of Operational Research*, *206*(3), 528–539. https://doi.org/10.1016/j.ejor.2010.02.032

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLOS ONE*, *14*(11), e0224365. https://doi.org/10.1371/journal.pone.0224365

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2020). Applying Machine Learning to Kinematic and Eye Movement Features of a Movement Imitation Task to Predict Autism Diagnosis. *Scientific Reports*, *10*(1), 8346. https://doi.org/10.1038/s41598-020-65384-4

van der Kruk, E., van der Helm, F. C. T., Veeger, H. E. J., & Schwab, A. L. (2018). Power in sports: A literature review on the application, assumptions, and terminology of mechanical power in sport research. *Journal of Biomechanics*, *79*, 1–14. https://doi.org/10.1016/j.jbiomech.2018.08.031

van der Worp, H., Vrielink, J. W., & Bredeweg, S. W. (2016). Do runners who suffer injuries have higher vertical ground reaction forces than those who remain injury-free? A systematic review and meta-analysis. *British Journal of Sports Medicine*, *50*(8), 450–457. https://doi.org/10.1136/bjsports-2015-094924

van Emmerik, R. E. A., Ducharme, S. W., Amado, A. C., & Hamill, J. (2016). Comparing dynamical systems concepts and techniques for biomechanical analysis. *Journal of Sport and Health Science*, *5*(1), 3–13. https://doi.org/10.1016/j.jshs.2016.01.013

van Ingen Schenau, G. J. (1989). From rotation to translation: Constraints on multi-joint movements and the unique action of bi-articular muscles. *Human Movement Science*, *8*(4), 301–337. https://doi.org/10.1016/0167-9457(89)90037-7

van Soest, A. J., & Bobbert, M. F. (1993). The contribution of muscle properties in the control of explosive movements. *Biological Cybernetics*, *69*(3), 195–204. https://doi.org/10.1007/BF00198959

van Soest, A. J., Bobbert, M. F., & Van Ingen Schenau, G. J. (1994). A control strategy for the execution of explosive movements from varying starting positions. *Journal of Neurophysiology*, *71*(4), 1390–1402. https://doi.org/10.1152/jn.1994.71.4.1390

Vanrenterghem, J., De Clercq, D., & Cleven, P. V. (2001). Necessary precautions in measuring correct vertical jumping height by means of force plate measurements. *Ergonomics*, *44*(8), 814–818. https://doi.org/10.1080/00140130118100

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, *7*(1), 91. https://doi.org/10.1186/1471-2105-7-91

Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, *180*, 68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061

Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2017). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage*, *145*, 166–179. https://doi.org/10.1016/j.neuroimage.2016.10.038

Veltink, P. H., Bussmann, HansB. J., de Vries, W., Martens, WimL. J., & Van Lummel, R. C. (1996). Detection of static and dynamic activities using uniaxial accelerometers. *IEEE Transactions on Rehabilitation Engineering*, *4*(4), 375–385. https://doi.org/10.1109/86.547939

Vig, J. R. (2000). Quartz Crystal Resonators and Oscillators. In *U.S. Army Communications—Electronic Command* (Vol. 1, Issue January).

Viitasalo, J. T. (1985). Measurement of the force-velocity characteristics for sportsmen in field conditions. In D. A. Winter, R. W. Norman, R. P. Well, K. C. Hayes, & A. E. Patla (Eds.), *International Series on Biomechanics* (pp. 96–101).

Wang, J., Chen, Y., Hao, S., Peng, X., & Hu, L. (2019). Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, *119*, 3–11. https://doi.org/10.1016/j.patrec.2018.02.010

Wang, Z., & Oates, T. (2015). Imaging Time-Series to Improve Classification and Imputation. *ArXiv:1506.00327 [Cs, Stat]*. http://arxiv.org/abs/1506.00327

Wang, Z., Zoghi, M., Hutter, F., Matheson, D., & De Freitas, N. (2013). Bayesian Optimization in High Dimensions via Random Embeddings. *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, 1778–1784.

Warmenhoven, J., Bargary, N., Liebl, D., Harrison, A. J., Robinson, M. A., Gunning, E., & Hooker, G. (2021). PCA of waveforms and functional PCA: A primer for biomechanics. *Journal of Biomechanics*, *116*, 110106. https://doi.org/10.1016/j.jbiomech.2020.110106

Warmenhoven, J., Cobley, S., Draper, C., Harrison, A. J., Bargary, N., & Smith, R. (2017a). Bivariate functional principal components analysis: Considerations for use with multivariate movement signatures in sports biomechanics. *Sports Biomechanics*, *18*(1), 10–27. https://doi.org/10.1080/14763141.2017.1384050

Warmenhoven, J., Cobley, S., Draper, C., Harrison, A. J., Bargary, N., & Smith, R. (2017b). Considerations for the use of functional principal components analysis in sports biomechanics: Examples from on-water rowing. *Sports Biomechanics*, *18*(3), 317–341. https://doi.org/10.1080/14763141.2017.1392594

Warmenhoven, J., Cobley, S., Draper, C., Harrison, A. J., Bargary, N., & Smith, R. (2017c). Assessment of propulsive pin force and oar angle time-series using functional data analysis in on-water rowing. *Scandinavian Journal of Medicine & Science in Sports*, *27*(12), 1688–1696. https://doi.org/10.1111/sms.12871

Warmenhoven, J., Cobley, S., Draper, C., Harrison, A. J., Bargary, N., & Smith, R. (2018). How gender and boat-side affect shape characteristics of force–angle profiles in single sculling: Insights from functional data analysis. *Journal of Science and Medicine in Sport*, *21*(5), 533–537. https://doi.org/10.1016/j.jsams.2017.08.010

Warmenhoven, J., Harrison, A. J., Robinson, M. A., Vanrenterghem, J., Bargary, N., Smith, R., Cobley, S., Draper, C., Donnelly, C., & Pataky, T. (2018). A force profile analysis comparison between functional data analysis, statistical parametric mapping and statistical non-parametric mapping in on-water single sculling. *Journal of Science and Medicine in Sport*, *21*(10), 1100–1105. https://doi.org/10.1016/j.jsams.2018.03.009

Wen, Q., Sun, L., Yang, F., Song, X., Gao, J., Wang, X., & Xu, H. (2021). Time Series Data Augmentation for Deep Learning: A Survey. *ArXiv:2002.12478 [Cs, Eess, Stat]*. http://arxiv.org/abs/2002.12478

Wiatrak, M., Albrecht, S. V., & Nystrom, A. (2020). Stabilizing Generative Adversarial Networks: A Survey. *ArXiv:1910.00927 [Cs]*. http://arxiv.org/abs/1910.00927

Wilson, J., Hutter, F., & Deisenroth, M. (2018). Maximizing acquisition functions for Bayesian optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 31* (pp. 9884–9895). Curran Associates, Inc. http://papers.nips.cc/paper/8194-maximizing-acquisition-functions-for-bayesian-optimization.pdf

Winter, E. M. (2005). JUMPING: POWER OR IMPULSE?: *Medicine & Science in Sports & Exercise*, *37*(3), 523. https://doi.org/10.1249/01.MSS.0000155703.50713.26

Wisloff, U. (2004). Strong correlation of maximal squat strength with sprint performance and vertical jump height in elite soccer players. *British Journal of Sports Medicine*, *38*(3), 285–288. https://doi.org/10.1136/bjsm.2002.002071

Wouda, F. J., Giuberti, M., Bellusci, G., Maartens, E., Reenalda, J., van Beijnum, B.-J. F., & Veltink, P. H. (2018). Estimation of Vertical Ground Reaction Forces and Sagittal Knee Kinematics During Running Using Three Inertial Sensors. *Frontiers in Physiology*, *9*, 218. https://doi.org/10.3389/fphys.2018.00218

Xu, Q.-S., & Liang, Y.-Z. (2001). Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, *56*(1), 1–11. https://doi.org/10.1016/S0169-7439(00)00122-2

Xu, Q.-S., Liang, Y.-Z., & Du, Y.-P. (2004). Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration. *Journal of Chemometrics*, *18*(2), 112–120. https://doi.org/10.1002/cem.858

Yang, D., Tang, J., Huang, Y., Xu, C., Li, J., Hu, L., Shen, G., Liang, C.-J. M., & Liu, H. (2017). TennisMaster: An IMU-based online serve performance evaluation system. *Proceedings of the 8th Augmented Human International Conference on - AH '17*, 1–8. https://doi.org/10.1145/3041164.3041186

Yang, Y. (2007). Consistency of cross validation for comparing regression procedures. *The Annals of Statistics*, *35*(6), 2450–2473. https://doi.org/10.1214/009053607000000514

Yoon, J., Jarrett, D., & van der Schaar, M. (2019). Time-series Generative Adversarial Networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 32). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/c9efe5f26cd17ba6216bbe2a7d26d490-Paper.pdf

Young, W., McLean, B., & Ardagna, J. (1995). Relationship between strength qualities and sprinting performance. *The Journal of Sports Medicine and Physical Fitness*, *35*(1), 13–19.

Young, W., Newton, R., Doyle, T., Chapman, D., Cormack, S., Stewart, C., & Dawson, B. (2005). Physiological and anthropometric characteristics of starters and non-starters and playing positions in elite Australian Rules football: A case study. *Journal of Science and Medicine in Sport*, *8*(3), 333–345. https://doi.org/10.1016/S1440-2440(05)80044-1

Yun, X., & Bachmann, E. R. (2006). Design, Implementation, and Experimental Results of a Quaternion-Based Kalman Filter for Human Body Motion Tracking. *IEEE Transactions on Robotics*, *22*(6), 1216–1227. https://doi.org/10.1109/TRO.2006.886270

Zago, Sforza, Dolci, Tarabini, & Galli. (2019). Use of Machine Learning and Wearable Sensors to Predict Energetics and Kinematics of Cutting Maneuvers. *Sensors*, *19*(14), 3094. https://doi.org/10.3390/s19143094

Zajac, F. E. (1993). Muscle coordination of movement: A perspective. *Journal of Biomechanics*, *26*, 109–124. https://doi.org/10.1016/0021-9290(93)90083-Q

Zhang, P. (1993). Model Selection Via Multifold Cross Validation. *The Annals of Statistics*, *21*(1). https://doi.org/10.1214/aos/1176349027

Zhang, Y., & Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, *187*(1), 95–112. https://doi.org/10.1016/j.jeconom.2015.02.006

Ziegert, J. C., & Lewis, J. L. (1979). The Effect of Soft Tissue on Measurements of Vibrational Bone Motion by Skin-Mounted Accelerometers. *Journal of Biomechanical Engineering*, *101*(3), 218–220. https://doi.org/10.1115/1.3426248