Australian & New Zealand Journal of Statistics

Aust. N. Z. J. Stat. **66**(1), 2024, 48–76 doi: 10.1111/anzs.12406

A novel response model and target selection method with applications to marketing

Y. Cai*

School of Management, Swansea University

Summary

Response models used in marketing are not always constructed for later marketing optimisation, which often results in unsatisfactory results in target selection for future marketing activities. To solve this problem, we develop a new binary response model and a new marketing target selection method. The proposed model can predict multiple propensity scores per customer through customer-specific propensity score distributions, which is not possible with existing response models, filling a gap in the literature. The target selection method can determine the best propensity scores from those predicted by the proposed model and use them to select customers for further marketing activities. Our simulation results and application to real marketing data confirm that the performance of the proposed model in target selection is significantly better than that of the existing models, including some popular machine learning methods, which indicate that our method can be very useful in practice.

Key words: marketing; propensity score; quantile function; response model; target selection.

1. Introduction

Statistical response models have been used in many disciplines, such as marketing. Many response models have been developed and used in marketing research. In addition to the probit and logit models, Alvarez & Brehm (1995) proposed a heterogeneous choice model to deal with possible heterogeneous issue in the data. Manski (1975, 1985) developed the binary quantile regression (BQR) model, which is a semi-parametric model, and Bult (1993) also discussed its advantages and limitations in marketing. Horowitz (1992) and Kordas (2006) proposed smoothed BQR models. Rossi, McCulloch & Allenby (1996) developed a random coefficient selection model to deal with the heterogeneity in observable demographics and applied the model to purchase history data. Train (2002) gave a more comprehensive discussion of this topic.

Moreover, an aggregate advertising response model based on consumer population dynamics was proposed by Wang *et al.* (2013), while Li & Ansari (2014) proposed a Bayesian semi-parametric approach for endogeneity and heterogeneity in choice models. Bruno, Cebollada & Chintagunta (2018) developed a new model and used it to deal with the intra-household heterogeneity in customer brand choice behaviour. Kappe, Blank &

*Author to whom correspondence should be addressed.
Department of Accounting and Finance, School of Management, Swansea University, Swansea SA1 8EN, UK. e-mail: y.cai@swansea.ac.uk

DeSarbo (2018) developed a random coefficients mixture hidden Markov model for flexible patterns of unobserved heterogeneity in both the state-dependent and transition parameters.

When using these response models in marketing, we first estimate the models by using a training data set, which contains customer information, including customer responses to specific product or service offers obtained in past or pilot marketing activities. Then, we use the estimated models to predict a propensity score for each customer in a test data set, which contains customer information that has not been used in the model estimation, where the propensity score indicates the likelihood of the customer becoming a potential buyer. Finally, the customers in the test set are divided into groups according to their propensity scores, which are arranged in a decreasing order. A group of customers (e.g. the top 10% of the customers) can then be selected for future marketing activities because they have a higher propensity score than others.

The response models used in marketing can perform well in target selection when the predicted propensity scores are unbiased. However, when the predicted propensity scores are biased, these response models' performance in target selection may not be satisfactory. This problem may occur, for example, when the response models are incorrectly specified and/or when data are imbalanced. In such cases, the propensity scores predicted by these models may be biased towards, for example non-buyers. As pointed out by Ling & Li (1998), the propensity scores may be skewed within a narrow range in one end, and thus, errors in propensity score estimation will affect the ranking more easily in target selection. Branco, Torgo & Ribeiro (2016) further pointed out that when data are imbalanced, a response model in marketing can perform well in prediction (as measured by the total percentage of correct predictions), but it may be unsatisfactory in target selection, as the predicted propensity scores may be biased towards non-buyers, resulting in worse target selection results.

It is seen that optimising the subsequent utilisation of the estimated response models may be necessary if they are employed to directly or indirectly assist marketing managers in making marketing-mix decisions. For example, we may need to use the estimated models to identify the optimal propensity scores that can be used to determine the best customer group that is most likely to be potential buyers of a company's new product or service.

However, the existing response models used in marketing are not always constructed for later marketing optimisation (Albers 2012), and the existing models do not allow us to predict multiple propensity scores for each customer. Therefore, we have no room to determine whether the propensity scores predicted by the models are the 'best' for target selection in marketing.

It is important to use optimal propensity scores in marketing because targeting using optimal propensity scores enables marketers to efficiently allocate resources, improve campaign effectiveness, reduce costs, and enhance customer satisfaction. It is also important to use optimal propensity scores in other subject areas, including psychology, political science, economics, finance and sociology, because using optimal propensity scores allows researchers in these fields to better address selection bias, assess the impact of policies and analyse observational data to make causal inferences.

However, to the best of our knowledge, there is currently no response model that predicts multiple propensity scores for each individual, allowing us to determine the optimal propensity score. It is this gap in the literature that motivates us to develop a new response model and related approach to target selection for marketing research in this paper.

The contribution of this work to the related literature is twofold. On the one hand, we contribute to the statistical literature by developing a novel binary response model that can be used to solve practical problems in many disciplines. Unlike existing response models, the proposed model explicitly estimates the entire propensity score distribution. This will play a vital role in later marketing optimisation because it will allow the proposed model to assign multiple propensity scores to each customer, and these propensity scores can cover a wide range of propensity score distribution, allowing us to identify optimal propensity scores for marketing. It is worth noting that this model can also be applied to other business problems including, for example, fraud detection and insurance/risk management.

On the other hand, we contribute to the marketing research by developing a target selection method in order to facilitate the use of the new model in marketing. Our target selection method includes two steps. The first step is to determine the optimal propensity score that should be used for selecting targets, and the second step is to use the identified propensity score to perform target selection in marketing. Since the best propensity score can be used for target selection, our model can deal with some of the problems caused by model specification errors and/or imbalanced data, as shown later in this paper.

Our results confirm that not all propensity scores can produce satisfactory results in target selection. This explains why if the propensity score estimated from a given model is biased, then the model's performance in target selection may be unsatisfactory. Our results also show that the proposed model outperforms baseline response models as well as some popular machine learning methods used in target selection. The main reason is that our method can use the 'best' propensity score for target selection, while the existing methods may not.

In Section 2, we briefly discuss the benchmark response models that are closely related to the model proposed in this paper, and in Section 3, we discuss the proposed response model and its estimation. In Section 4, we present the method for target selection. Simulation results are discussed in Section 5, while in Section 6, we discuss the application of the proposed method to the bank marketing data, and compare the results with those obtained from the benchmark response models and some machine learning methods commonly used in marketing. Section 7 concludes. Appendix A gives the proofs of three theorems, Appendix B and Appendix C give the prior density function and the MCMC method for parameter estimation, respectively, and Appendix D presents results of another simulation study which confirms that the proposed estimation method performs well, and the convergence of the method does not depend on the strength of the prior information on the parameters.

2. A brief review of the baseline response models

The binary logistic regression (BLR) model is one of the popular response models used for target selection in marketing. Many researchers use it as a benchmark model, see for example Cui, Wong & Wan (2012) and Zahavi & Levin (1997).

Generally, a binary regression model is defined by

$$\begin{aligned} &y_i^* = \eta(\mathbf{x}_i, \boldsymbol{\alpha}) + \varepsilon_i, \\ &y_i = 1 \text{ if } y_i^* > 0; \quad y_i = 0 \text{ otherwise}, \end{aligned} \tag{1}$$

where $i = 1, \ldots, n$ and n is the sample size, $\boldsymbol{\alpha}$ is a vector of parameters, $\mathbf{x}_i = (x_{1i}, \ldots, x_{ki})$ is the observed value of k predictors for customer i, and y_i is the observed response of customer i. Moreover, y_i^* is an unobserved continuous variable that may represent, for example, customers' psychological feelings about a new product, and ε_is are assumed to be iid random variables.

Let $F(\cdot)$ be the distribution function of ε_i and $\eta_i = \eta(\mathbf{x}_i, \boldsymbol{\alpha})$. Then, the propensity score of customer i can be estimated by the response probability of the customer, which is given by $\mu_i = \Pr(y_i = 1|\eta_i) = \Pr(y_i^* > 0|\eta_i) = \Pr(\varepsilon_i > -\eta_i) = 1 - F(-\eta_i)$. If $F(\cdot)$ is symmetric about zero, then we also have $\mu_i = F(\eta_i)$. When $F(\eta_i)$ is the logistic distribution function defined by $F(\eta_i) = e^{\eta_i}/(1 + e^{\eta_i})$, model (1) becomes the BLR model. In this case, the conditional mean of y_i^*, that is η_i, is used to estimate the propensity score μ_i. It is worth mentioning that, unless otherwise stated, in this paper we assume that ε_i follows the logistic distribution.

Note that model (1) cannot handle heterogeneous issues that may exist in the data. In other words, when the variance of ε_i is not a constant, model (1) is not useful. Yatchew & Griliches (1985) pointed out that in the presence of heterogeneity, the estimation of model parameters will be inconsistent and inefficient. In order to solve the problem of heterogeneity, Alvarez & Brehm (1995) proposed a heterogeneous choice (HC) model:

$$\begin{aligned} y_i^* &= \eta(\mathbf{x}_i, \boldsymbol{\alpha}_1) + e^{\mathbf{z}_i'\boldsymbol{\alpha}_2}\varepsilon_i, \\ y_i &= 1 \text{ if } y_i^* > 0, \quad y_i = 0; \text{ otherwise,} \end{aligned} \tag{2}$$

where $\mathbf{z}_i$ is also a vector of predictors that may be the same as $\mathbf{x}_i$, and $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ are vectors of the model parameters. So, model (2) tries to deal with the heterogeneous problem by using the term $e^{\mathbf{z}_i'\boldsymbol{\alpha}_2}$. However, Achen (2002) pointed out that model (2) is equivalent to the following model

$$\begin{aligned} y_i^* &= \eta(\mathbf{x}_i, \boldsymbol{\alpha}_1)/e^{\mathbf{z}_i'\boldsymbol{\alpha}_2} + \varepsilon_i, \\ y_i &= 1 \text{ if } y_i^* > 0, \quad y_i = 0; \text{ otherwise,} \end{aligned}$$

and there is no way to distinguish between them. This implies that model (2) is in fact a non-linear homogeneous binary response model. Hence, the predicted propensity score is given by $\mu_i = e^a/(1 + e^a)$, where $a = \eta(\mathbf{x}_i, \boldsymbol{\alpha}_1)/e^{\mathbf{z}_i'\boldsymbol{\alpha}_2}$.

The BQR model developed by Manski (1975, 1985), Horowitz (1992) and Kordas (2006) is a semi-parametric model because they do not assume a distribution for the error term, and they can be estimated by using the methods based on Manski's (1975, 1985) maximum score function and its variants. The benefits and limitations of semi-parametric models are also discussed by Bult (1993). Recently, Benoit & Van den Poel (2012) developed a Bayesian approach to the BQR model by using the asymmetric Laplace distribution for ε_i, denoted by ALD($\psi = 0, \sigma = 1, \tau$), where σ and ψ are the scale and location of the distribution, respectively, and $\tau \in (0, 1)$. Hence the model can be expressed by

$$\begin{aligned} y_i^* &= \eta(\mathbf{x}_i, \boldsymbol{\alpha}_1) + \varepsilon_i, \\ y_i &= 1 \text{ if } y_i^* > 0; \quad y_i = 0 \text{ otherwise,} \\ \varepsilon_i &\sim \text{ALD}(\psi = 0, \sigma = 1, \tau). \end{aligned} \tag{3}$$

As the τ quantile of ALD is zero, $\eta_i = \eta(\mathbf{x}_i, \boldsymbol{\alpha}_1)$ is in fact the conditional τ quantile of y_i^*. Following Hashem *et al.* (2016), the propensity score can be estimated by $\mu_i = 1 - \text{ALD}(\psi = -\eta(\mathbf{x}_i, \boldsymbol{\alpha}_1), \sigma = 1, \tau = 0.5)$, which corresponds to the conditional median of y_i^*.

We will use the BLR, HC and BQR models as benchmark response models because they are closely related to our model. This is discussed more at the end of Section 3.1.

3. Proposed model and parameter estimation

The construction of the existing response models discussed above determines that only one propensity score can be assigned to each customer. Hence, if the predicted propensity scores are biased, the errors in these scores will affect the ranking of the customers, leading to unsatisfactory results in target selection.

In order to overcome the limitations of existing models, in this section, we first develop a new response model by explicitly estimating the distribution of propensity scores, and then we develop a parameter estimation method.

3.1. Proposed model

Recall that, for the BLR model, the propensity score is given by $\mu_i = \Pr(y_i = 1|\eta_i) = e^{\eta_i}/(1 + e^{\eta_i})$. If we regard μ_i as a random variable, then η_i is also a random variable. Hence, if we know the distribution of η_i, then we also know the distribution of μ_i. To treat η_i as a random variable, we first propose the following model:

$$\begin{aligned} y_i^* &= \eta_i + \varepsilon_i, \\ \eta_i &= h_1(\boldsymbol{\alpha}, \mathbf{x}_i) + h_2(\boldsymbol{\beta}, \mathbf{x}_i)\xi_i, \\ y_i &= 1 \text{ if } y_i^* > 0, \quad y_i = 0; \text{ otherwise,} \end{aligned} \tag{4}$$

where $h_2(\boldsymbol{\beta}, \mathbf{x}_i) > 0$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are parameter vectors to be estimated, both ε_i and ξ_i are iid random variables and ε_i follows the logistic distribution. We further assume ε_i and ξ_i are independent.

Model (4) says that $h_1(\boldsymbol{\alpha}, \mathbf{x}_i)$ and $h_2(\boldsymbol{\beta}, \mathbf{x}_i)$ are the location and scale of the distribution of η_i respectively, where the predictors involved in h_1 and h_2 may be different, but to simplify the notation used in the paper, we set them equal. It is seen that the term $h_2(\boldsymbol{\beta}, \mathbf{x}_i)$ can also be used to deal with some heterogeneity in the data.

The second step of our model building process is to determine the distribution of ξ_i in model (4). For target selection, we need the distribution of η_i to capture different characteristics of data, such as the characteristics related to the centre, skewness and tails of the data. This is important because, for example, when data are imbalanced, the distribution of the data is skewed. Compared with the centre, the tail of the distribution may contain more useful information that can be used for target selection.

The work of Fournier *et al.* (2007) suggests that the generalised lambda distribution (GLD) can be a very good candidate for our model because this distribution is so flexible that many standard distributions, such as normal, Weibull, log-normal, t-, skewed t- and F-distributions, and many other distributions can be accurately approximated by this distribution. This shows that the GLD can handle many data structures that these standard distributions may not be able to handle alone. Hence, using the GLD for ξ_i will make the model more robust to model specification errors.

It is worth noting that the GLD is only explicitly defined by its quantile function, which can be expressed by $Q_{gld}(\tau) = a + bQ(\tau, \boldsymbol{\gamma})$, where a and b are the location and the scale of the distribution, and $Q(\tau, \boldsymbol{\gamma})$ is defined by

$$Q(\tau, \boldsymbol{\gamma}) = \frac{\tau^{\gamma_1} - 1}{\gamma_1} - \frac{(1-\tau)^{\gamma_2} - 1}{\gamma_2}, \quad \gamma_1 < 0,\ \gamma_2 < 0, \tag{5}$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$ and $\tau \in (0, 1)$. In fact, $Q(\tau, \boldsymbol{\gamma})$ defined by (5) is a special case of GLD, with location 0 and scale 1. So we let ξ_i follow the GLD defined by (5). However, as the GLD is only explicitly defined by its quantile function, we need to use the quantile function to rewrite model (4), which gives

$$\begin{aligned} & y_i^* = \eta_i + \varepsilon_i, \\ & Q_{\eta_i}(\tau \mid \boldsymbol{\theta}, \mathbf{x}_i) = h_1(\boldsymbol{\alpha}, \mathbf{x}_i) + h_2(\boldsymbol{\beta}, \mathbf{x}_i) Q(\tau, \boldsymbol{\gamma}), \\ & y_i = 1 \text{ if } y_i^* > 0, \quad y_i = 0;\ \text{otherwise}, \end{aligned} \tag{6}$$

where $\tau \in (0, 1)$ and $Q(\tau, \boldsymbol{\gamma})$ represents the quantile function of ξ_i and $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$. In this paper, we refer to model (6) as the quantile function response (QFR) model.

Theorem 1. *Let $\eta_i = h_1(\boldsymbol{\alpha}, \mathbf{x}_i) + h_2(\boldsymbol{\beta}, \mathbf{x}_i)\xi_i$, and let $Q(\tau, \boldsymbol{\gamma})$ be the quantile function of ξ_i. Then the conditional quantile function of η_i is given by $Q_{\eta_i}(\tau \mid \boldsymbol{\theta}, \mathbf{x}_i) = h_1(\boldsymbol{\alpha}, \mathbf{x}_i) + h_2(\boldsymbol{\beta}, \mathbf{x}_i) Q(\tau, \boldsymbol{\gamma})$. Moreover, model (6) implies that the conditional quantile function of the propensity score μ_i is given by*

$$Q_{\mu_i}(\tau \mid \boldsymbol{\theta}, \mathbf{x}_i) = e^{Q_{\eta_i}(\tau \mid \boldsymbol{\theta}, \mathbf{x}_i)} \Big/ \left(1 + e^{Q_{\eta_i}(\tau \mid \boldsymbol{\theta}, \mathbf{x}_i)}\right). \tag{7}$$

See Appendix A for a proof. Theorem 1 shows that model (6) actually defines the entire distribution of η_i through its quantile function. It further shows that the distribution of propensity scores can also be obtained easily through (7). Moreover, due to the monotone relation between η_i and μ_i, the τ quantile of η_i corresponds to the τ quantile of the propensity score μ_i for any $\tau \in (0, 1)$. This shows another advantage of using the quantile function of η_i in the model. We will take advantage of this when we develop the target selection method later in the paper.

In the following, we focus on model (6), where $Q(\tau, \boldsymbol{\gamma})$ is defined by (5). Hence, $Q_{\eta_i}(\tau|\boldsymbol{\theta}, \mathbf{x}_i)$ defines the GLD with location $h_1(\boldsymbol{\alpha}, \mathbf{x}_i)$ and scale $h_2(\boldsymbol{\beta}, \mathbf{x}_i)$. Moreover, γ_1 and γ_2 control its left and right tails respectively. If $\gamma_1 = \gamma_2$, then the distribution of η_i is symmetric, otherwise it is skewed. Furthermore, as Gilchrist (2000) pointed out, γ_1 and γ_2 determine not only the skewness but also the relative weights of the tails. This means that the skewness is modelled as a result of tail shape rather than as an independent feature. This is one of the features that other standard distributions may not have.

The final step of our model building process is to specify a functional form for $h_1(\boldsymbol{\alpha}, \mathbf{x}_i)$ and $h_2(\boldsymbol{\beta}, \mathbf{x}_i)$ respectively. In order to compare with the benchmark response models, we use the form of a linear function for $h_1(\boldsymbol{\alpha}, \mathbf{x}_i)$. We use the form of a quadratic function for $h_2(\boldsymbol{\beta}, \mathbf{x}_i)$ because estimating generalised non-linear regression models that contain exponential forms can be difficult due to large sampling errors (see, McCullagh & Nelder 1989). Specifically, in this paper, we let

$$h_1(\boldsymbol{\alpha}, \mathbf{x}_i) = \alpha_0 + \alpha_1 x_{1i} + \cdots + \alpha_k x_{ki}, \quad h_2(\boldsymbol{\beta}, \mathbf{x}_i) = \beta_0 + \beta_1 x_{1i}^2 + \cdots + \beta_k x_{ki}^2, \tag{8}$$

where $\beta_0 > 0$ and $\beta_j \geq 0, j = 1, \ldots, k$, which guarantee that $h_2(\boldsymbol{\beta}, \mathbf{x}_i) > 0$.

Theorem 2. *Consider model (6). If $Q(\tau, \boldsymbol{\gamma})$ is given by (5) and $h_1(\boldsymbol{\alpha}, \mathbf{x}_i)$ and $h_2(\boldsymbol{\beta}, \mathbf{x}_i)$ are defined by (8), then model (6) is well defined on the parameter space $\overline{\Omega} = \{(\alpha_j, \beta_j, \gamma_v) : -\infty < \alpha_j < \infty,\ \beta_0 > 0, \beta_j \geq 0\,(j \neq 0),\ \gamma_v < 0,$ all possible $j, v\}$.*

See Appendix A for a proof. Therefore, once the quantile function $Q_{\eta_i}(\tau|\boldsymbol{\theta}, \mathbf{x}_i)$ is available, we know the entire distribution of η_i and μ_i. As the distribution of η_i and μ_i defined by the model is very flexible, many features of the data can be captured by the distribution, such as features related to centre, dispersion, skewness, tail and so on. The captured features can then be regarded as potential propensity scores for target selection. We will discuss this issue in Section 4.

The relations between the QFR model (6) and the BLR, HC and BQR models are discussed below. If $h_2(\boldsymbol{\beta}, \mathbf{x}_i) = 0$, then the QFR model becomes the BLR model. The HC model tries to deal with the heterogeneity through the unobserved y_i^*, although the model is equivalent to a non-linear homogeneous model, while the QFR model can deal with heterogeneity through η_i. The QFR model uses the quantile *function* approach to response modelling (see, e.g. Gilchrist 2000) and hence its parameters do not depend on τ, while the BQR model uses the quantile *regression* approach to response modelling (see, e.g. Koenker 2005) and hence its parameters depend on τ. In addition, the QFR model can assign multiple propensity scores to each customer, while the BLR, HC and BQR models can only assign a unique propensity score to each customer. Finally, all response models, namely QFR, BLR, HC and BQR models, do not consider specific data structures, such as panel data, time series data or data with specific heterogeneous structures. Hence, they are good benchmark response models for our research.

3.2. Parameter estimation

In order to facilitate the use of the proposed model in practice, we now discuss how to estimate the model parameters. In this paper, we develop an estimation method for model (6), where $Q(\tau, \boldsymbol{\gamma})$, $h_1(\boldsymbol{\alpha}, \mathbf{x}_i)$ and $h_2(\boldsymbol{\beta}, \mathbf{x}_i)$ are defined by (5) and (8) respectively. However, this estimation method can be easily extended to other formulae for $Q(\tau, \boldsymbol{\gamma})$, $h_1(\boldsymbol{\alpha}, \mathbf{x}_i)$ and $h_2(\boldsymbol{\beta}, \mathbf{x}_i)$.

3.2.1. Posterior density function

First recall that in our model, the propensity score μ_i is a random variable whose distribution is determined by its quantile function $Q_{\mu_i}(\tau|\boldsymbol{\theta}, \mathbf{x}_i)$. So a random sample of μ_i, also denoted by μ_i in order to simplify the notation used in this paper, is given by $\mu_i = Q_{\mu_i}(\tau_i|\boldsymbol{\theta}, \mathbf{x}_i)$, where τ_i is a random sample of the uniform distribution between 0 and 1.

Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)$, $\mathrm{y} = (y_1, \ldots, y_n)$ and $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$. Then given $\mathbf{x}$, the likelihood of $\mathbf{y}$ can be expressed by $L(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\theta}, \mathbf{x}) = \prod_{i=1}^n \mu_i^{y_i}(1 - \mu_i)^{(1-y_i)}$. However, since μ_i depends on τ_i and τ_i is unobservable, the MLE method is not convenient for the parameter estimation. Thus we consider a Bayesian approach to parameter estimation, which requires us to derive the posterior distribution of $\boldsymbol{\theta}$, $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$.

Theorem 3. *Let $\pi(\boldsymbol{\mu},\boldsymbol{\tau},\boldsymbol{\theta}|\mathbf{x},\mathbf{y})$ and $\pi_0(\boldsymbol{\theta}|\mathbf{x})$ be the posterior and prior density functions of $(\boldsymbol{\mu},\boldsymbol{\tau},\boldsymbol{\theta})$ and $\boldsymbol{\theta}$ respectively. Then*

$$\pi(\boldsymbol{\mu},\boldsymbol{\tau},\boldsymbol{\theta}|\mathbf{x},y) \propto \left\{\prod_{i=1}^{n}\mu_i^{y_i}(1-\mu_i)^{(1-y_i)}\pi_i(\mu_i|\tau_i,\boldsymbol{\theta},\mathbf{x}_i)\right\}\pi_0(\boldsymbol{\theta}|\,\mathbf{x}), \tag{9}$$

where $\pi_i(\mu_i|\tau_i,\boldsymbol{\theta},\mathbf{x}_i) = \frac{\left\{1+e^{Q_{\eta_i}(\tau_i|\boldsymbol{\theta},x_i)}\right\}^2}{e^{Q_{\eta_i}(\tau_i|\boldsymbol{\theta},\mathbf{x}_i)}\frac{dQ(\tau_i,\boldsymbol{\gamma})}{d\tau}h_2(\boldsymbol{\beta},\mathbf{x}_i)}$.

Furthermore, let $\Omega = \Omega_1 \times \Omega_2 \times \Omega_3$, where $\boldsymbol{\theta} \in \Omega_1 = \{(\alpha_j,\beta_j,\gamma_v)|\alpha_j \in [-M,M],\ \beta_j \in [\epsilon,M],\ \gamma_v \in [-M,-\epsilon]$, for all possible $j,v\}$, $\boldsymbol{\mu} \in \Omega_2 = (0,1)^n$ and $\boldsymbol{\tau} \in \Omega_3 = [\epsilon,1-\epsilon]^n$, in which M and ϵ are two fixed positive real numbers. Suppose $\pi_0(\boldsymbol{\theta}|\mathbf{x})$ is well defined on Ω_1. Then the posterior density function $\pi(\boldsymbol{\mu},\boldsymbol{\tau},\boldsymbol{\theta}|\mathbf{x},y)$ defined by (9) is well defined on Ω in the sense that $\int_\Omega \pi(\boldsymbol{\mu},\boldsymbol{\tau},\boldsymbol{\theta}|\mathbf{x},\mathbf{y})d\boldsymbol{\mu}d\boldsymbol{\tau}d\boldsymbol{\theta} < \infty$.

See Appendix A for a proof. In this paper, we let $M = 10^{20}$ and $\epsilon = 10^{-20}$. Hence, the difference between Ω_1 and $\overline{\Omega}$ (see Theorem 2) is negligible, thus ensuring that important parameter regions in $\overline{\Omega}$ will not be missed. Theorem 3 shows that the posterior density function is well defined on Ω, but it is very complicated. Hence, the Markov Chain Monte Carlo (MCMC) method is suitable for parameter estimation.

3.2.2. MCMC method

It follows from Theorem 3 that we now need to specify a prior density function $\pi_0(\boldsymbol{\theta}|\mathbf{x})$ so that it is well defined on Ω_1. To simplify the calculation, we let $\pi_0(\boldsymbol{\theta}|\mathbf{x}) = \pi(\boldsymbol{\alpha})\pi(\boldsymbol{\beta})\pi(\boldsymbol{\gamma})$, where α_j follows a truncated normal distribution on $[-M,M]$, and β_j and $-\lambda_v$ follow a truncated log-normal distribution on $[\epsilon,M]$ respectively. Appendix B provides the detailed formula of the prior density function.

The basic idea of a MCMC method is to generate a sequence of values in the parameter space Ω so that this sequence of values forms a Markov Chain whose equilibrium distribution is the posterior distribution of the parameters. See Brooks (1998) for details. To achieve this, we use the Metropolis-Hastings algorithm in which a candidate parameter value is simulated from a chosen distribution and this proposed value is accepted as the next in the sequence with a known probability, see Geyer (2011). The detailed steps of our MCMC method are given in Appendix C.

Therefore, after a burn-in period, posterior samples can be collected from the Markov Chain generated by the MCMC method. The model parameters can be estimated by using the average of the posterior samples. We denote the estimated parameters by $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$. Then, the conditional quantile function of η_i and μ_i can be estimated by $\hat{Q}_{\eta_i}(\tau|\hat{\boldsymbol{\theta}},\mathbf{x}_i) = h_1(\hat{\boldsymbol{\alpha}},\mathbf{x}_i) + h_2\left(\hat{\boldsymbol{\beta}},\mathbf{x}_i\right)Q(\tau,\hat{\boldsymbol{\gamma}})$ and $\hat{Q}_{\mu_i}(\tau|\hat{\boldsymbol{\theta}},\mathbf{x}_i) = e^{\hat{Q}_{\eta_i}(\tau|\hat{\boldsymbol{\theta}},x_i)}/\{1+e^{\hat{Q}_{\eta_i}(\tau|\hat{\boldsymbol{\theta}},\mathbf{x}_i)}\}$ respectively. Using the distribution of η_i or μ_i, we can now assign multiple propensity scores to each customer for target selection.

4. Target selection method

4.1. Multiple propensity scores

Recall that if we use an existing model, we can only assign a unique propensity score to each customer. If, for example, the model is not specified correctly or the data are imbalanced, the predicted propensity scores may be biased towards non-buyers, resulting in large errors in target selection.

However, for the QFR model, we have the entire propensity score distribution, defined by $\hat{Q}_{\mu_i}(\tau|\hat{\boldsymbol{\theta}}, \mathbf{x}_i)$. Therefore, we can use, for example, the mean, median, quantiles or other information about the distribution of μ_i as the estimated propensity scores for each customer. It can be seen that each customer now has multiple propensity scores, and different propensity scores contain information about different parts (or characteristics) of the μ_i distribution.

It is worth noting that which propensity score is more suitable for target selection depends on the data structure. For example, if the data are balanced, a propensity score containing information about the centre of the μ_i distribution may be more appropriate, but when the data are imbalanced, a propensity score containing information about the tail of the distribution may be more appropriate. Therefore, a wide range of propensity scores can avoid losing important information about the distribution of μ_i and enable us to determine the 'best' propensity score for target selection.

It is worth emphasising again that the relation between η_i and the corresponding μ_i is monotonic. This suggests that the ranking of customers based on μ_i is the same as the ranking of customers based on η_i, thus leading to the same results in target selection. Therefore, in order to simplify the calculation, in the rest of this paper, we will use η_i to define the propensity scores for target selection.

Let η_{ij} be the jth propensity score of customer i, where $j = 1, \ldots, J$, and J is the total number of propensity scores that are assigned to customer i. Let $C_j = \{\eta_{ij}, i = 1, \ldots, n\}$. Then C_j contains the jth propensity score of all customers.

In order to ensure that η_{ij} can cover a wide distribution range of η_i, we can let η_{ij} be the τ_j quantile of η_i, that is, $\eta_{ij} = Q_{\eta_i}(\tau_j|\boldsymbol{\theta}, \mathbf{x}_i)$, where $0 < \tau_1 < \cdots < \tau_J < 1$ cover a wide range between 0 and 1. As an example, we can define first seven propensity scores by letting $\tau_j \in \{0.01, 0.05, 0.25, 0.5, 0.75, 0.95, 0.99\}$, we can define the eighth propensity score by letting $\eta_{i8} = \eta_{i7} - \eta_{i1}$. Therefore, these propensity scores reflect the main characteristics of the η_i distribution or μ_i distribution. More specifically, the tail of the distribution is captured by $\eta_{i1}, \eta_{i2}, \eta_{i6}$ and η_{i7}, the central location is captured by the η_{i3}, η_{i4} and η_{i5}, and the dispersion is captured by η_{i8}. It is worth noting that other information about the distribution of η_i may also be used to define propensity scores, but in this paper, we will mainly use the propensity scores defined above.

4.2. Target selection

It is worth noting that in direct marketing, the lift is usually the most important criterion for assessing the performance of a model in target selection (see e.g. Bhattacharya 1999, Vuk & Curk 2006, and Cui, Wong & Wan 2012). The lift is defined by the ratio of the percentage of true positive responses in a specific group of customers identified by a model to the percentage of responses in the group identified by the random model. For example, a model with a lift of 2 in a group of 10% customers is said to be twice as good as the random model in that group. Thus, we can rank customers in an ordered list: customers with higher propensity scores are at the top of the list. Then, we can calculate the lift in the top 10% of customers, and then calculate the lift in the top 20% of customers, and so on, until we reach 100% of customers. Using lifts across these groups is helpful for comparing the performance of different models.

For our model, after assigning multiple propensity scores to each customer, we need to determine which score should be used in target selection. Recall that we have used the training set to estimate the model. We now still need to use the training set to determine the best propensity score for target selection. The main steps of our target selection method are given below.

(i) Given j, rank customers in the training set to obtain an ordered list: customers with higher jth propensity scores are at the top of the list.
(ii) Use the ordered customer list to define 10 customer groups, denoted by $A_{\ell j}$, where $\ell = 1, \ldots, 10$, and $A_{\ell j}$ contains the top $10\ell\%$ customers.
(iii) Calculate the lift in each group $A_{\ell j}$, denoted by $u_{\ell j}$.
(iv) Calculate the average value of the lifts, denoted by $\bar{u}_j$.
(v) Repeat the above steps for all $j = 1, \ldots, J$.
(vi) Let $\bar{u}_{j^*} = \max\{\bar{u}_j, j = 1, \ldots, J\}$. Then the j^*th propensity score is the propensity score that should be used for target selection.
(vii) Calculate the j^*th propensity score of all customers in the test set, rank the customers and select a group of customers for future marketing activities.

Note that $\max\{\bar{u}_j, j = 1, \ldots, J\}$ may not be unique. For example, we may have $\bar{u}_2 = \bar{u}_6 = \max\{\bar{u}_j, j = 1, \ldots, J\}$. In this case, both propensity scores may be used for target selection. Users can determine which one to use in practice. It is also worth noting that if the training and test data sets are not random samples of the same data set, then the proposed target selection method may not work, because in this case, the two data sets may have different data structures. To ensure that the training and test data sets have the same structure, various sampling methods may be used. For example, the simple random sampling method, stratified random sampling method or combinations of several random sampling methods can be used.

In the above method, we used the average of the lifts in the 10 percentage groups to determine which propensity score should be used in target selection. Our results show that this simple optimisation criterion works well. However, it is worth noting that we can also use other criteria to determine the best propensity score to use. For example, instead of the average lift, we may use a median lift, we may use the area under the lift curve, or we may combine multiple propensity scores with the financial cost of marketing (see e.g. Bult & Wansbeek 1995). Therefore, the comparison between different optimisation standards needs to be studied in the future.

5. Simulation study

We now discuss the simulation results in order to gain some in-depth understanding of the performance of the proposed model in target selection. The purpose of this simulation study is to confirm that, in the presence of model specification errors, the performance of the proposed model in marketing is better than the benchmark response models. This is important because in practice, we often face the problems caused by model specification errors.

We first simulated x_i from $\mathrm{N}(0,1)$ for $i=1,\dots,500$, and then we simulated η_i from $\mathrm{N}(m_i,\sigma_i)$, where $m_i = 0.2+0.5x_i^2+0.25\sin(x_i)$, $\sigma_i = 0.2+0.11e^{x_i}$ for $i=1,\dots,250$, and $m_i=-0.5-0.5x_i+0.25x_i^2$, $\sigma_i=0.5$ for $i=251,\dots,500$. The μ_i values were calculated by $\mu_i = e^{\eta_i}/(1+e^{\eta_i})$. Finally, we let $y_i=1$ with probability μ_i and $y_i=0$ otherwise.

The above procedure was repeated 200 times, resulting in 200 independent data sets, the first of which is shown in Figure 1.

It is seen that the data structure is quite complicated and the dependence between variable x and η is not linear. We fitted the following four models to each of the data sets.

QFR model: $y_i^* = \eta_i + \varepsilon_i$, where

$$Q_{\eta_i}(\tau\,|\,\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}) = \alpha_0+\alpha_1x_i+(\beta_0+\beta_1x_i^2)\left(\frac{\tau^{\gamma_1}-1}{\gamma_1}-\frac{(1-\tau)^{\gamma_2}-1}{\gamma_2}\right)$$

and $Q_{\mu_i}(\tau\,|\,\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma})$ is given by (7).

BLR model: $y_i^* = \eta_i+\varepsilon_i$, where $\eta_i = c_0+c_1x_i$ and $\mu_i = e^{\eta_i}/(1+e^{\eta_i})$.

BQR model: $y_i^* = \eta_i+\varepsilon_i$, where $\eta_i = d_0+d_1x_i$ and $\mu_i = 1-\mathrm{ALD}(\psi=-(d_0+d_1x_i),\sigma=1,\tau=0.5)$.

HC model: $y_i^* = \eta_i+\sigma_i\varepsilon_i$, where $\eta_i=e_0+e_1x_i$, $\sigma_i = e^{e_2x_i}$ and $\mu_i = e^{\eta_i/\sigma_i}/(1+e^{\eta_i/\sigma_i})$.

Therefore, all four response models are incorrectly specified. We now consider the performance of these models in prediction and in target selection. When predicting discrete variables, commonly used metrics are the confusion matrix (Kohavi & Provost 1998) and the area under the receiver operating characteristic (ROC) curve (Fawcett 2006), denoted by AUC. The larger the AUC, the higher the predictive power of the model. To generate

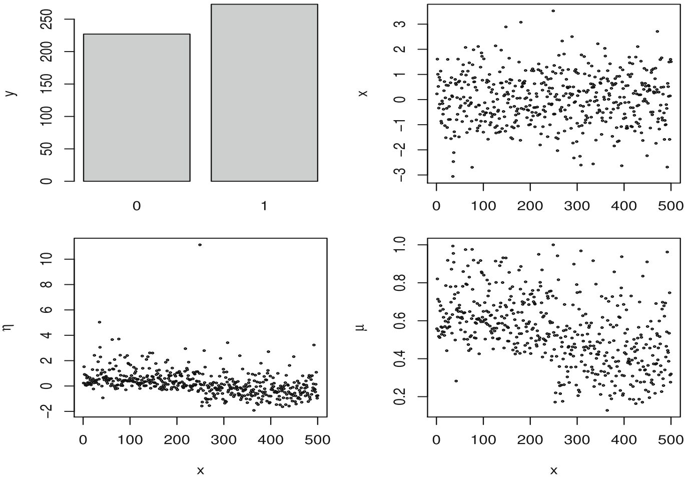

Figure 1. Plots of the first simulated data set in the simulation study.

an ROC curve, we need to (i) calculate the propensity score μ_i of the customers in a data set, (ii) select a sequence of score thresholds $0 < c_1 < \cdots < c_K < 1$, (iii) for each c_k, let $y_i = 1$ if $\mu_i > c_k$, and $y_i = 0$ otherwise, and calculate the total percentage (denoted by p_k) of correct positive predictions and (iv) plot p_k against c_k to obtain the ROC curve for a model. It can be seen that the higher the ROC curve, the larger the AUC, and therefore the better the prediction performance of the response model. It is worth noting that in this paper, we use the median of μ_i (corresponding to the median of η_i) to predict the response variable.

Figure 2 shows the plot of the AUCs in 200 simulations. It is seen that the AUC values obtained from our QFR model are larger than those obtained from the SLR, HC and BQR models. In fact, we also used the one-tailed t-test to check the significance of the difference between the average AUCs obtained from different models. The results show that, at the 1% significance level, the QFR model has the highest predictive power, the HC model has

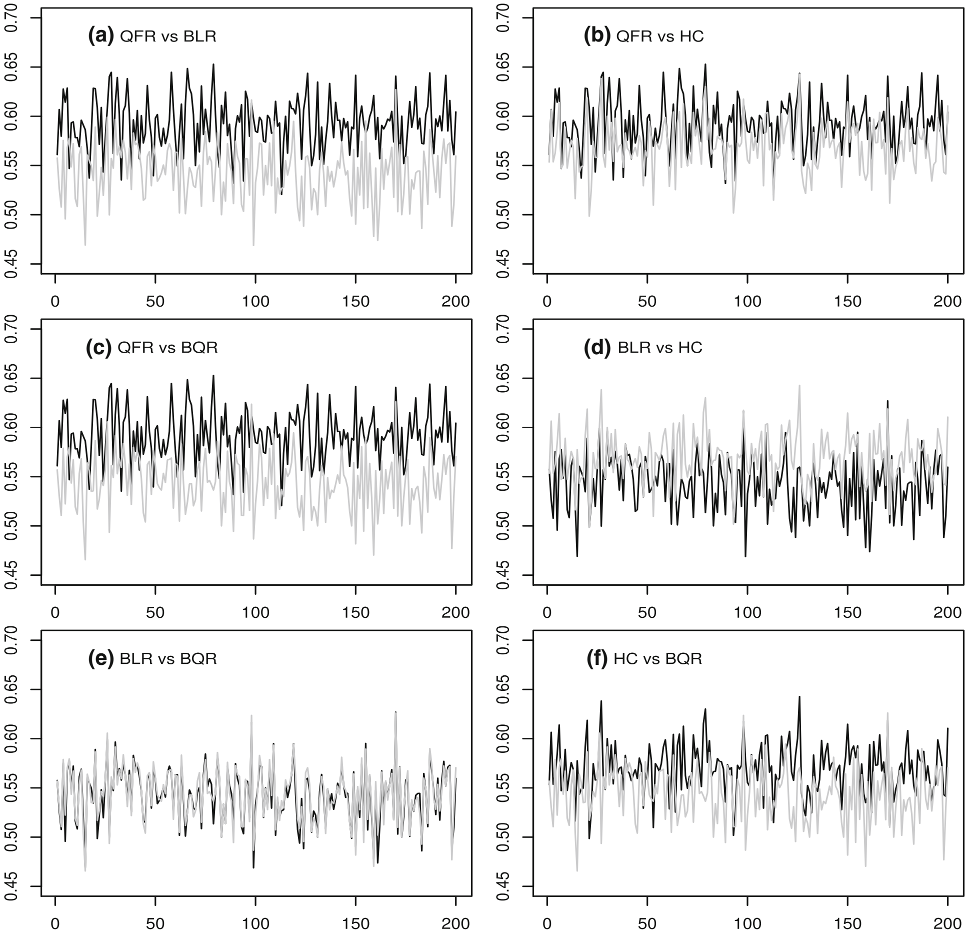

Figure 2. AUC plots for the fitted models. (a)–(c) Black curves for QFR model and grey curves for BLR, HC and BQR models. (d, e) Black curves for BLR model and grey curves for HC and BQR models. (f) Black curve for HC model and grey curve for BQR model.

the second highest predictive power, and the BQR and BLR models have similar predictive power.

Now, we consider target selection. We used the first simulated data set as the training set, and simulated another 100,000 data from the same model to give the test set. For the QFR model, we use the median of μ_i and η_{ij} as propensity scores for customer i, where η_{ij} is the τ_j quantile of $\eta_i, j = 1, \ldots, 5$ and $\tau_j \in \{0.05, 0.25, 0.5, 0.75, 0.95\}$. Therefore, in this simulation study, we assigned a total of six propensity scores to each customer. By using the training set and the target selection method discussed in Section 4.2, we calculated the lift in each group, the results of which are given in Table 1, where μ represents the median of μ_i. So, we can use Table 1 to determine which propensity score should be used in target selection.

It is seen that the lifts obtained based on the propensity scores defined by the median of μ_i and the 50% quantile of η_i are the same. This is not surprising, because the median of μ_i corresponds to the 50% quantile of η_i (i.e. the median of η_i), they should produce the same result in target selection. Comparing the average lift, Table 1 also shows that the average lift based on these two propensity scores is the largest. Therefore, we can use the 50% quantile of η_i or the median of μ_i for target selection.

Next, we check whether the best result of target selection on the test set still corresponds to the best propensity score determined above. Note that in reality, the test set does not contain values of y_i. However, since we use the simulated data, we know the values of y_i in the test set and we can use them to check whether the best propensity score identified above is still the best on the test set.

Table 2 shows the lifts obtained from all propensity scores predicted by the QFR model on the test set. It can be seen that the results on the test set are very similar to those on the training set, with the best results again corresponding to the 50% quantile of η_i or the median of μ_i. This is what we should expect, because both the training set and the test set are simulated from the same model, and hence they have the same data structure.

We also checked performance of the benchmark response models in target selection and included the results in Table 2. It can be seen that the overall performance of our model in target selection is also better than that of the benchmark response models. Note that all three benchmark response models have the same lifts because they are accurate to two decimal places.

Table 1. Lifts for propensity score determination.

Group	μ	η_{i1} 5%	η_{i2} 25%	η_{i3} 50%	η_{i4} 75%	η_{i5} 95%
10%	1.40	1.01	1.11	1.40	1.29	1.26
20%	1.16	0.98	0.98	1.16	1.18	1.18
30%	1.10	0.98	0.98	1.10	1.09	1.08
40%	1.09	0.92	0.94	1.09	1.08	1.06
50%	1.09	0.94	0.94	1.09	1.04	1.05
60%	1.05	0.97	0.97	1.05	1.04	1.04
70%	1.03	0.96	0.96	1.03	1.02	1.01
80%	1.00	0.96	0.95	1.00	1.01	1.00
90%	1.01	0.96	0.96	1.01	1.00	1.00
100%	1.00	1.00	1.00	1.00	1.00	1.00
Average	1.09	0.97	0.98	1.09	1.07	1.07

Table 2. Lifts for test set in target selection.

		η_{i1}	η_{i2}	η_{i3}	η_{i4}	η_{i5}			
Group	μ	5%	25%	50%	75%	95%	BLR	HC	BQR
10%	1.42	0.88	0.88	1.42	1.38	1.34	1.43	1.43	1.43
20%	1.30	0.87	0.87	1.30	1.27	1.27	1.27	1.27	1.27
30%	1.22	0.87	0.87	1.22	1.21	1.21	1.17	1.17	1.17
40%	1.16	0.88	0.88	1.16	1.16	1.15	1.11	1.11	1.11
50%	1.12	0.89	0.89	1.12	1.12	1.11	1.06	1.06	1.06
60%	1.09	0.90	0.90	1.09	1.08	1.08	1.03	1.03	1.03
70%	1.06	0.91	0.91	1.06	1.06	1.06	1.00	1.00	1.00
80%	1.03	0.93	0.93	1.03	1.03	1.03	0.99	0.99	0.99
90%	1.01	0.96	0.96	1.01	1.01	1.01	0.99	0.99	0.99
100%	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Average	**1.14**	0.91	0.91	**1.14**	1.13	1.13	1.10	1.10	1.10

We further conducted another simulation study in order to confirm that the proposed estimation method performs well and the convergence of the method does not depend on the strength of the prior information on the parameters. Indeed, good results were also obtained. See Appendix D for details.

6. A marketing application

In this section, we compare the performance of the four response models in terms of prediction and target selection. We also compare the performance of these response models in target selection with some popular machine learning methods commonly used in marketing.

6.1. The data

The data considered in this application were collected from marketing campaigns conducted by a Portuguese banking institution in the period from May 2008 to November 2010. Although the original data are not available, Moro, Laureano & Cortez (2011) used data mining techniques and provided a subset of the data, which contains 16 attributes, one output variable and 45211 instances, of which 5289 were successful. Hence the success rate was 11.7%.

The variables given in the data set are defined as follows. The response variable is Y, where $Y = 1$ represents that the customer subscribed a term deposit and $Y = 0$ otherwise. The 16 predictors are age (V_1), job type (V_2), marital status (V_3), education level (V_4), credit in default (V_5), average yearly balance in euros (V_6), housing loan (V_7), personal loan (V_8), contact communication type (V_9), last contact day of the month (V_{10}), last contact month of year (V_{11}), last contact duration (V_{12}), number of contacts performed during this campaign (V_{13}), number of days that passed by after the client was last contacted from a previous campaign (V_{14}), number of contacts performed before this campaign and for this client (V_{15}) and the outcome of the previous marketing campaign (V_{16}).

We divided the data set into two random subsets, namely the training set and the test set. The training set consists of 690 customers, and the test set contains the rest of the data. The sample size of 690 was randomly selected between 600 and 800 using the random

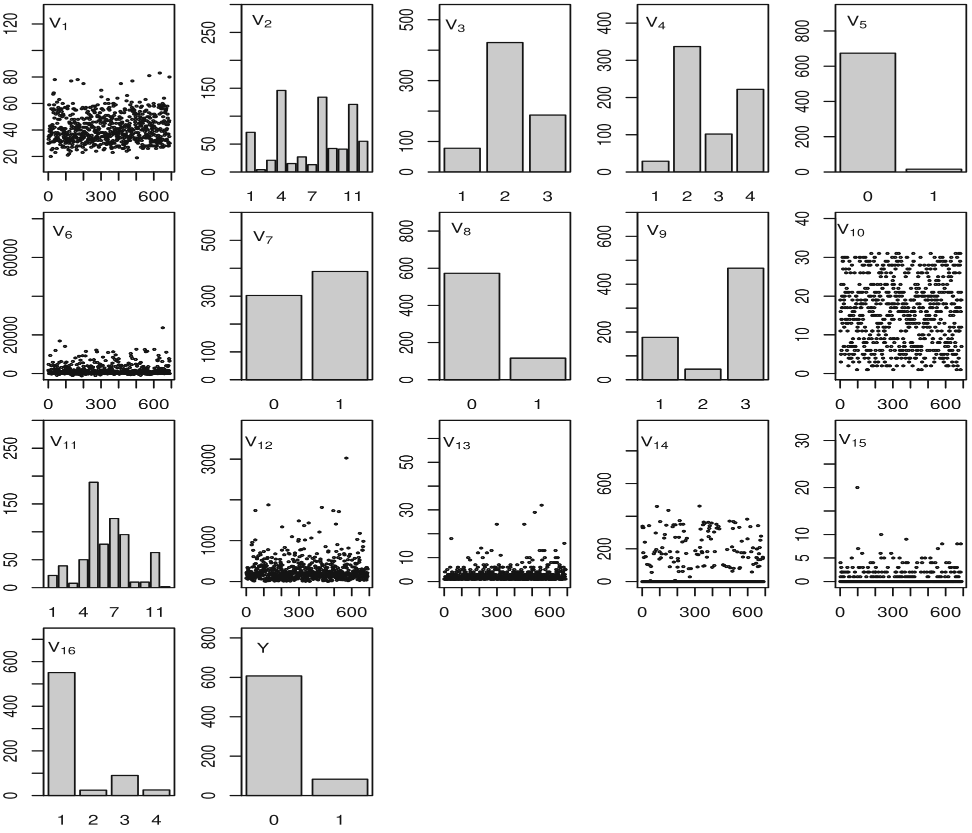

Figure 3. Plots of the observed bank marketing data in the training set.

number generator in the statistical software R. As we do not know the specific structure of the entire data set, we used the simple random sampling method to obtain the training set. We deliberately keep the sample size of the training set small, because in practice, we often need to conduct pilot marketing activities for a small group of customers before we conduct a full scale marketing campaign. Therefore, response models that perform well on smaller data sets are very important in practice.

We will use the training set to estimate the models, and use the test set to examine the performance of models in target selection. For our model, we will also use the training set to determine the propensity score that should be used in target selection. Plots of the data in the training set are shown in Figure 3. It is seen that the data are imbalanced. In fact, the success rate in the training set is 13.6%, very similar to the entire data set.

6.2. Estimated response models

We estimated the four response models using the training set. For the QFR model defined by (6), (5) and (8), a Markov Chain was run for 5×10^5 steps, whose plot (not shown to save space) suggests that the Markov Chain converged after a burn-in period of the first 10^4 steps. The three benchmark response models were estimated using the

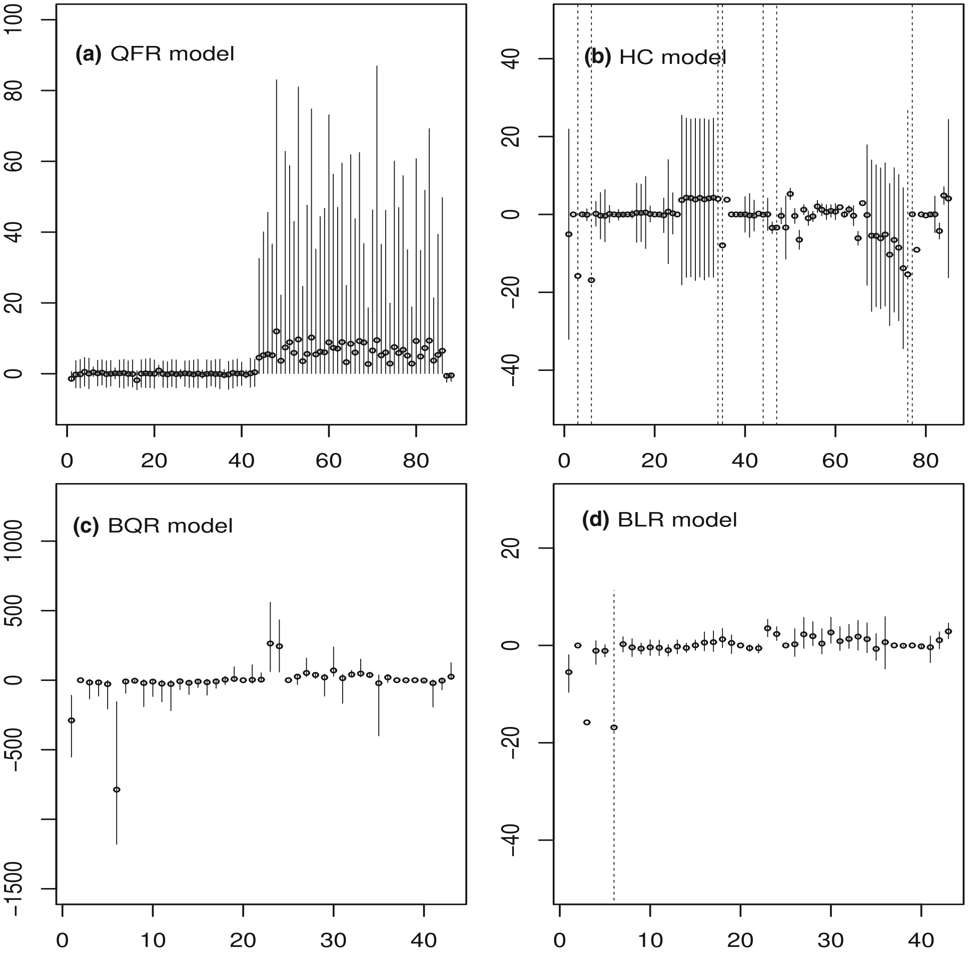

Figure 4. The parameters (points) of the four response models estimated using the bank marketing data. The vertical line segment shows the 95% credible/confidence interval. The vertical dashed line indicates that the confidence interval is outside the range of the graph.

statistical software R. Figure 4 summarises the estimated parameters (dots), where each vertical line segment corresponds to the 95% credible interval or confidence interval of a model parameter, and each vertical dashed line indicates that the confidence interval is too wide to be displayed within the range of the graph.

6.3. Performance of the response models in prediction

We first consider in-sample prediction. We used the four fitted models to calculate the propensity score of customers in the training set and obtained the total percentage of correct positive predictions at each score threshold. Then, we show the ROC curves of the four models in Figure 5a, where the vertical line corresponds to the typical score threshold 0.5. It can be seen that for in-sample prediction, our model has better performance than the benchmark response models in the entire range of 0–1.

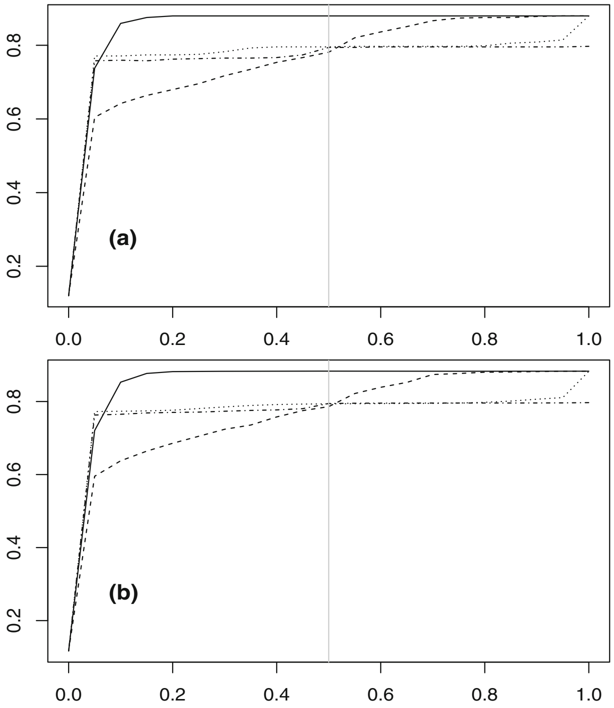

Figure 5. (a) ROC curves on the training set. (b) ROC curves on the test set. The vertical line corresponds to the score threshold 0.5. The continuous, dash-dotted, dotted and dashed ROC curves correspond to QFR, HC, BQR and BLR models respectively.

For the out-of-sample prediction, we check the average performance of the models. We obtained 100 independent random subsets from the test set, each of which contains 10,000 customers. Then, for each of these 100 subsets, we used the four fitted models to calculate the propensity scores and obtain the total percentage of correct positive predictions.

It is worth mentioning that we do not have any information about the data structure, and therefore, we use 100 independent random subsets of the test set for out-of-sample prediction, which will allow us to check how the models perform on different test sets that may have different unknown structures.

To make the plot clearer, we further calculated the average percentage of correct positive predictions at each score threshold and show the corresponding ROC curves in Figure 5b. Clearly, the performance of our model in prediction is also better than that of the benchmark response models. On the other hand, in the benchmark response models, when the score threshold is less than 0.5, the HC and BQR models are better than the BLR model, otherwise the BLR model is better.

6.4. Performance of the response models in target selection

A response model can perform well in prediction, but it may be unsatisfactory in target selection, especially in the case of imbalanced data (Branco, Torgo & Ribeiro 2016). We now examine the performance of the models in target selection. First, we consider our model.

We use the method discussed in Section 4.1 to define the propensity scores in this study. Specifically, for $j \leq 7$, let η_{ij} be the jth propensity score of customer i, where η_{ij} is the τ_j quantile of η_i, and $\tau_j \in \{0.01, 0.05, 0.25, 0.5, 0.75, 0.95, 0.99\}$. For $j = 8$, let $\eta_{i8} = \eta_{i7} - \eta_{i1}$. Let C_j contain the jth propensity score of all customers in the training set.

After ranking customers in the training set according to the propensity scores contained in C_j, we calculated the lift in each percentage group. Table 3 shows all the lifts obtained using the training set. It is worth noting that Table 3 does not contain the results corresponding to the propensity scores defined by μ_i, because they are the same as the results given in Table 3.

It is seen that the maximum average lift corresponds to $j = 5, 6, 7, 8$. According to our target selection method, the last four propensity scores are the best propensity scores, and we can use any of them to select targets from the test set.

Figure 6 shows the lifts in Table 3, where the grey and darker curves correspond to $j \leq 4$ and $j \geq 5$ respectively. Obviously, the grey curves are much lower than the

Table 3. Lifts for propensity score determination.

Groups	η_{i1} 1%	η_{i2} 5%	η_{i3} 25%	η_{i4} 50%	η_{i5} 75%	η_{i6} 95%	η_{i7} 99%	η_{i8} 99%-1%
10%	0.98	0.98	0.98	0.61	**1.22**	**1.22**	**1.22**	**1.22**
20%	1.04	1.11	1.04	0.86	**1.54**	**1.54**	**1.54**	**1.54**
30%	1.07	1.07	1.07	0.70	**1.60**	**1.60**	**1.56**	**1.56**
40%	0.96	0.96	0.93	0.62	**1.42**	**1.42**	**1.42**	**1.42**
50%	0.84	0.84	0.84	0.67	**1.16**	**1.18**	**1.18**	**1.21**
60%	0.76	0.76	0.76	0.70	**1.07**	**1.07**	**1.07**	**1.07**
70%	0.74	0.74	0.74	0.72	**1.02**	**1.01**	**1.01**	**1.01**
80%	0.85	0.87	0.85	0.82	**0.96**	**0.97**	**0.97**	**0.96**
90%	0.98	0.98	0.98	0.95	**1.03**	**1.03**	**1.03**	**1.03**
100%	1.00	1.00	1.00	1.00	**1.00**	**1.00**	**1.00**	**1.00**
Average	0.92	0.93	0.92	0.76	**1.20**	**1.20**	**1.20**	**1.20**

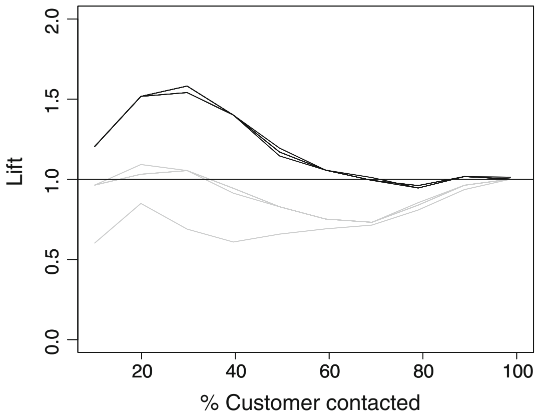

Figure 6. The lift chart for propensity score determination. The grey and darker curves correspond to $j \leq 4$ and $j \geq 5$ respectively. The horizontal line represents the baseline.

horizontal line, which means that if the first four propensity scores are used in target selection, the performance of the model in target selection will be much worse than the random model. On the other hand, the dark curves are mainly located above the horizontal line, especially for the first few percentage groups. All these results indicate that if the last four propensity scores are used, the model will produce similar results in target selection, and all these results will be much better than the results obtained from the random model. Therefore, any of the last four propensity scores can be used for target selection.

Now, given $j \geq 5$, we can calculate the jth propensity score of the customers in the test set, and rank the customers in the test set according to their jth propensity scores. Then, we can select a certain percentage of customers for future marketing activities. Please note that in reality, until the completion of the marketing activities, the performance of the model on the test set for target selection is not known. However, in this study, since the data were collected in the past bank marketing activities, we can use a lift table or lift chart to check the performance of the model on the test set.

More specifically, for comparison purposes, we calculated all eight propensity scores for each customer in the test set, as well as the lift in each percentage group. We present the results in Table 4 and Figure 7.

As expected, Table 4 clearly shows that the last four propensity scores have produced very similar results in target selection, and these results are also much better than those obtained by using the first four propensity scores. It is worth noting that Table 4 also confirms that, when the data are imbalanced, the propensity scores corresponding to the centre of the distribution (50% quantile of η_i) may not be suitable for target selection.

Figure 7 further shows that the lift curves corresponding to the last four propensity scores are much higher than other lift curves, so confirming again that the best propensity scores identified on the training set do produce the best results in target selection on the test set.

We further used the benchmark response models to calculate the propensity scores of customers in the test set and the lift in each percentage group. The results can also be found from Table 4 and Figure 7. Obviously, in terms of target selection, the performance of all benchmark response models is much worse than our model. In fact, they are even worse than the random model in target selection.

Table 4. Lifts obtained from different models on the test set.

	η_{i1}	η_{i2}	η_{i3}	η_{i4}	η_{i5}	η_{i6}	η_{i7}	η_{i8}			
Groups	1%	5%	25%	50%	75%	95%	99%	99%-1%	BLR	BQR	HC
10%	0.91	0.90	0.87	0.56	1.83	1.84	1.85	1.86	0.36	0.32	0.44
20%	0.82	0.82	0.79	0.51	1.84	1.85	1.86	1.86	0.43	0.38	1.02
30%	0.79	0.79	0.76	0.50	1.76	1.78	1.78	1.79	0.54	0.50	0.88
40%	0.71	0.71	0.69	0.51	1.50	1.51	1.51	1.53	0.68	0.65	0.76
50%	0.67	0.67	0.66	0.53	1.30	1.31	1.32	1.33	0.78	0.75	0.71
60%	0.65	0.65	0.64	0.56	1.17	1.18	1.18	1.19	0.87	0.80	0.68
70%	0.66	0.66	0.66	0.62	1.08	1.08	1.09	1.09	0.92	0.80	0.64
80%	0.78	0.78	0.78	0.75	1.03	1.04	1.04	1.04	0.91	0.77	0.70
90%	0.90	0.90	0.90	0.86	1.00	1.01	1.01	1.01	0.96	0.77	0.73
100%	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Average	0.79	0.79	0.77	0.64	1.35	1.36	1.36	1.37	0.75	0.67	0.76

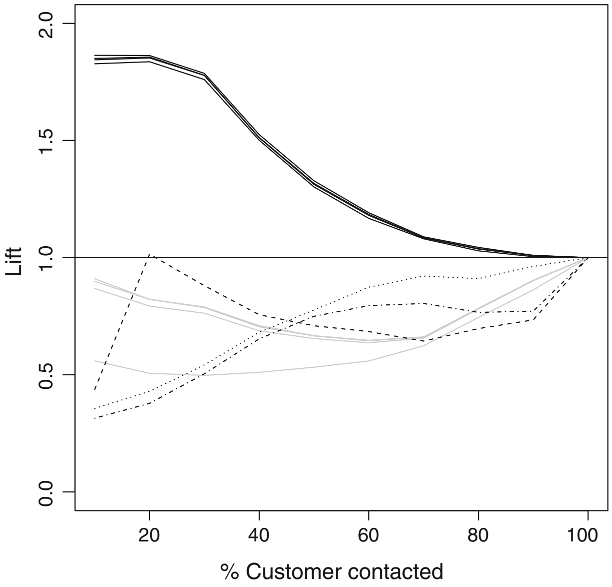

Figure 7. The lift chart for the test set: Grey and black curves correspond to the QFR model with $j \leq 4$ and $j \geq 5$ respectively. Dotted, dot-dashed and dashed curves correspond to BLR, BQR and HC models respectively. Horizontal line represents the baseline.

It is worth noting that in this application, the propensity scores obtained from the benchmark response models correspond to the centre of a skewed distribution, which is similar to our fourth propensity score (that is, the 50% quantile of η_i). Therefore, in this study, they are not as useful as the propensity scores corresponding to the tails of the η_i distribution in the target selection.

6.5. Performance of some machine learning methods in target selection

Now we further compare our methods with the following popular machine learning methods that are commonly used in marketing. (a) Decision tree (DT) (see, e.g. Hastie, Tibshirani & Friedman 2009), which is a machine learning method that recursively partitions data based on features to create a tree-like structure that predicts target outcomes by traversing the tree's branches according to feature values. (b) Classification and regression trees (CART) (see, e.g. Breiman *et al.* 1984), which is a DT-based approach for categorical (classification) and continuous (regression) target prediction, where data are divided into subsets at each node according to features, resulting in leaf nodes containing the predicted target value or class. (c) Random forests trees (RFT) (see, e.g. Breiman, 2001), which is an ensemble learning technique that combines multiple DTs, improves prediction accuracy by averaging their outputs and provides robust and reliable target selection by aggregating predictions from different individual trees. (d) K-nearest neighbour classification method (KNN) (see, e.g. Cover & Hart 1967), which assigns target labels to data points by considering the majority class of the k nearest neighbors in the feature space. (e) Quantile regression random forests (QRRF) (see, e.g. Meinshausen 2006), which is a hybrid approach that combines the ensemble power of random forests with quantile regression, allowing

Table 5. Lifts obtained from machine learning methods.

Groups	DT	CART	RFT	KNN	QRRF	NBC	SVM
10%	0.82	1.10	0.62	0.62	0.63	0.76	0.67
20%	0.97	1.10	0.78	0.78	0.77	0.80	0.80
30%	1.01	1.10	0.85	0.85	0.85	0.86	0.87
40%	1.03	1.10	0.89	0.89	0.89	0.91	0.91
50%	1.05	1.10	0.92	0.92	0.92	0.94	0.94
60%	1.05	1.10	0.94	0.94	0.94	0.95	0.96
70%	1.06	1.10	0.96	0.96	0.96	0.96	0.97
80%	1.05	1.10	0.98	0.98	0.98	0.98	0.98
90%	1.04	1.00	0.99	0.99	0.99	0.99	0.99
100%	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Average	1.01	1.08	0.89	0.89	0.89	0.92	0.91

prediction of different quantiles of the target variable, thus providing a more comprehensive understanding of the distribution of the data. (f) Naïve Bayes classification (NBC) method (see, e.g. Hastie, Tibshirani & Friedman 2009), which applies Bayes theorem and assumes feature independence to assign target classes to data points based on the probabilistic relationship between their features and class probabilities. (g) Support vector machine (SVM) classifier (see, e.g. Cortes & Vapnik 1995), which classifies data points into different classes by finding a hyperplane to maximise the separation between data points in feature space, enabling efficient target selection even in complex and high-dimensional data. It is worth noting that some of these methods also work well with imbalanced data.

We use each of these methods to perform the target selection and calculate the lift in each percentage group. The results are given in Table 5. Compared with the results obtained from the response models, we see that these machine learning methods outperform the BLR, BQR and HC methods, and the best machine learning method is the classification and regression trees method with an average lift of 1.08. However, these results are still much worse than the results obtained from our method.

7. Conclusion

We have developed a new response model that can explicitly estimate the entire propensity score distribution. Therefore, it can assign multiple propensity scores to each customer, which provides a way to fill the gap in the literature. To facilitate the use of the proposed model in marketing, we also developed a new target selection method. The target selection method can be used to identify the best propensity score from the propensity scores predicted by the proposed model, and use the identified propensity score to select targets for future marketing activities.

In this paper, we discussed our approach by using a special case of the QFR model (6), where $Q(\tau, \boldsymbol{\gamma})$ is given by (5) and $h_1(\boldsymbol{\alpha}, \mathbf{x}_i)$ and $h_2(\boldsymbol{\beta}, \mathbf{x}_i)$ are given by (8). In fact, the methodology developed in the paper can be easily extended to other QFR models defined by different $Q(\tau, \boldsymbol{\gamma})$, $h_1(\boldsymbol{\alpha}, \mathbf{x}_i)$ and $h_2(\boldsymbol{\beta}, \mathbf{x}_i)$ functions.

We have seen that the propensity scores identified by our target selection method can indeed produce much improved results in target selection compared to the benchmark response models and the machine learning methods commonly used in marketing. Our results also confirm that not all propensity scores can produce good results in target selection. This explains why the propensity scores predicted by existing response models may not provide satisfactory results in target selection.

It is seen that our model helps us understand and better capture complex patterns and dependencies in customer data, and more accurately predict customer behaviour and responses to marketing campaigns. With our enhanced target selection method, marketers can identify and target specific customer segments more effectively. On the other hand, it is also worth noting that the model we have developed can also be applied to other business problems including, for example, fraud detection and insurance/risk management.

Finally, our target selection method is based on the average lift to determine the optimal propensity score for target selection, which is limited. Therefore, future examination and comparison of many other criteria that may be helpful in determining the optimal propensity score for target selection will be required.

APPENDIX A

Proof of Theorem 1. As $Q(\tau, \boldsymbol{\gamma})$ is the quantile function of ξ_i, we have $\Pr\{\xi_i \le Q(\tau, \boldsymbol{\gamma})\} = \tau$ for any $\tau \in (0, 1)$. It follows from $\tau = \Pr\{\xi_i \le Q(\tau, \boldsymbol{\gamma})\} = \Pr\{h_1(\boldsymbol{\alpha}, \mathbf{x}_i) + h_2(\boldsymbol{\beta}, \mathbf{x}_i)\xi_i \le h_1(\boldsymbol{\alpha}, \mathbf{x}_i) + h_2(\boldsymbol{\beta}, \mathbf{x}_i)Q(\tau, \boldsymbol{\gamma})\} = \Pr\{\eta_i \le h_1(\boldsymbol{\alpha}, \mathbf{x}_i) + h_2(\boldsymbol{\beta}, \mathbf{x}_i)Q(\tau, \boldsymbol{\gamma})\}$ that the τ quantile of η_i is given by $h_1(\boldsymbol{\alpha}, \mathbf{x}_i) + h_2(\boldsymbol{\beta}, \mathbf{x}_i)Q(\tau, \boldsymbol{\gamma})$. As τ is an arbitrary real number between 0 and 1, we see that the quantile function of η_i is given by $Q_{\eta_i}(\tau \mid \boldsymbol{\theta}, \mathbf{x}_i) = h_1(\boldsymbol{\alpha}, \mathbf{x}_i) + h_2(\boldsymbol{\beta}, \mathbf{x}_i)Q(\tau, \boldsymbol{\gamma})$, which is the second equation in model (6) as required.

Note that $\mu_i = \Pr(y_i = 1|\eta_i) = e^{\eta_i}/(1 + e^{\eta_i})$ is a monotone function of η_i. Hence, given the quantile function $Q_{\eta_i}(\tau \mid \boldsymbol{\theta}, \mathbf{x}_i)$ of η_i, the quantile function of μ_i is $Q_{\mu_i}(\tau \mid \boldsymbol{\theta}, \mathbf{x}_i) = e^{Q_{\eta_i}(\tau \mid \boldsymbol{\theta}, \mathbf{x}_i)}/(1 + e^{Q_{\eta_i}(\tau \mid \boldsymbol{\theta}, \mathbf{x}_i)})$ as required.

Proof of Theorem 2. We need to show that two distinct parameter vectors, $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ and $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)$ cannot yield the same value of the maximised likelihood function.

First note that if for the two parameter vectors $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ and $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)$ such that

$$Q_{\eta_i}(\tau \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{x}_i) = Q_{\eta_i^*}(\tau \mid \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \mathbf{x}_i) \tag{A1}$$

holds for all $\mathbf{x}_i$ and $\tau \in (0, 1)$, then we have

$$\prod_{i=1}^{n} \mu_i^{y_i}(1 - \mu_i)^{1-y_i} = \prod_{i=1}^{n} (\mu_i^*)^{y_i}(1 - \mu_i^*)^{1-y_i},$$

where the quantile function of μ_i^* is given by

$$Q_{\mu_i^*}(\tau \mid \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \mathbf{x}_i) = e^{Q_{\eta_i^*}(\tau|\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \mathbf{x}_i)} \Big/ \left\{1 + e^{Q_{\eta_i^*}(\tau \mid \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \mathbf{x}_i)}\right\}.$$

Hence, we only need to show that if (A1) holds, then we must have $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$, $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ and $\boldsymbol{\gamma} = \boldsymbol{\gamma}^*$.

For (A1) to hold, we need, for all $\mathbf{x}_i$ and $\tau \in (0, 1)$,

$$\alpha_0 + \alpha_1 x_{1i} + \cdots + \alpha_k x_{ki} = \alpha_0^* + \alpha_1^* x_{1i} + \cdots + \alpha_k^* x_{ki}, \tag{A2}$$

and

$$\left(\beta_0 + \beta_1 x_{1i}^2 + \cdots + \beta_k x_{ki}^2\right)\left(\frac{\tau^{\gamma_1}-1}{\gamma_1} - \frac{(1-\tau)^{\gamma_2}-1}{\gamma_2}\right)$$
$$= \left(\beta_0^* + \beta_1^* x_{1i}^2 + \cdots + \beta_k^* x_{ki}^2\right)\left(\frac{\tau^{\gamma_1^*}-1}{\gamma_1^*} - \frac{(1-\tau)^{\gamma_2^*}-1}{\gamma_2^*}\right). \tag{A3}$$

It follow from (A2) that, for all $\mathbf{x}_i$

$$(\alpha_0 - \alpha_0^*) + (\alpha_1 - \alpha_1^*)x_{1i} + \cdots + (\alpha_k - \alpha_k^*)x_{ki} = 0,$$

Hence, we must have $\alpha_i = \alpha_i^*$ for $i = 0, \ldots, k$, that is $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$.

For (A3) to hold for all $\mathbf{x}_i$ and all $\tau \in (0, 1)$, we need

$$\beta_0 + \beta_1 x_{1i}^2 + \cdots + \beta_k x_{ki}^2 = \beta_0^* + \beta_1^* x_{1i}^2 + \cdots + \beta_k^* x_{ki}^2$$

for all $\mathbf{x}_i$, which means that $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ must hold; and

$$\frac{\tau^{\gamma_1}-1}{\gamma_1} - \frac{(1-\tau)^{\gamma_2}-1}{\gamma_2} = \frac{\tau^{\gamma_1^*}-1}{\gamma_1^*} - \frac{(1-\tau)^{\gamma_2^*}-1}{\gamma_2^*}$$

for all $\tau \in (0, 1)$. This means that two strictly monotone quantile functions are equal and the parameters do not depend on τ. Therefore, we must have $\gamma_1 = \gamma_1^*$ and $\gamma_2 = \gamma_2^*$, that is $\boldsymbol{\gamma} = \boldsymbol{\gamma}^*$ as required. This completes the proof.

Proof of Theorem 3. First, note that if we let $f(x)$ be the probability density function of a random variable X and $Q_x(\tau)$ its quantile function, then we have $f(x) = \{dQ_x(\tau)/d\tau\}^{-1}$. Hence

$$\pi(\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\theta} \,|\, \mathbf{x}, y) \propto L(\mathbf{y} \,|\, \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\theta}, \mathbf{x})\pi(\boldsymbol{\mu} \,|\, \boldsymbol{\tau}, \boldsymbol{\theta}, \mathbf{x})\pi(\boldsymbol{\tau} \,|\, \boldsymbol{\theta}, \mathbf{x})\pi_0(\boldsymbol{\theta} \,|\, \mathbf{x})$$
$$= \left\{\prod_{i=1}^{n} \mu_i^{y_i}(1-\mu_i)^{(1-y_i)}\pi_i(\mu_i \,|\, \tau_i, \boldsymbol{\theta}, \mathbf{x}_i)\pi_i(\tau_i \,|\, \boldsymbol{\theta}, \mathbf{x}_i)\right\}\pi_0(\boldsymbol{\theta} \,|\, \mathbf{x}),$$

where $\pi_i(\mu_i \,|\, \boldsymbol{\theta}, \mathbf{x}_i)$ is given by

$$\pi_i(\mu_i \,|\, \boldsymbol{\theta}, \mathbf{x}_i) = \left\{\frac{dQ_{\mu_i}(\tau \,|\, \boldsymbol{\theta}, \mathbf{x}_i)}{d\tau}\right\}_{\tau=\tau_i}^{-1}$$
$$= \left[\frac{e^{Q_{\eta_i}(\tau \,|\, \boldsymbol{\theta}, \mathbf{x}_i)}\frac{dQ(\tau, \boldsymbol{\gamma})}{d\tau}h_2(\boldsymbol{\beta}, \mathbf{x}_i)}{\left\{1 + e^{Q_{\eta_i}(\tau \,|\, \boldsymbol{\theta}, x_i)}\right\}^2}\right]_{\tau=\tau_i}^{-1} = \frac{\left\{1 + e^{Q_{\eta_i}(\tau_i \,|\, \boldsymbol{\theta}, x_i)}\right\}^2}{e^{Q_{\eta_i}(\tau_i \,|\, \boldsymbol{\theta}, \mathbf{x}_i)}\frac{dQ(\tau_i, \boldsymbol{\gamma})}{d\tau}h_2(\boldsymbol{\beta}, \mathbf{x}_i)},$$

and $\pi_i(\tau_i \,|\, \boldsymbol{\theta}, \mathbf{x}_i) = 1$ as τ_i is uniformly distributed on $(0, 1)$. This completes the proof.

Since $\mu_i \in (0, 1)$, we see that $\mu_i^{y_i}(1-\mu_i)^{(1-y_i)} \leq 1$. Moreover, for $\tau_i \in [\epsilon, 1-\epsilon]$, the density function $\pi_i(\mu_i \,|\, \boldsymbol{\theta}, \mathbf{x}_i)$ is continuous on the closed set $[\epsilon, 1-\epsilon]$, hence it is finite on $[\epsilon, 1-\epsilon]$. Therefore, there exists a constant, say $\tilde{M}$, such that $\prod_{i=1}^{n} \mu_i^{y_i}(1-\mu_i)^{(1-y_i)}\pi_i(\mu_i \,|\, \boldsymbol{\theta}, \mathbf{x}_i) \leq \tilde{M}$ on Ω.

Hence, it follows from $\int_{\Omega_1} \pi_0(\boldsymbol{\theta}\,|\,\mathbf{x})d\boldsymbol{\theta} < \infty$ that $\int_{\Omega} \pi(\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\theta}\,|\,\mathbf{x}, \mathbf{y})d\boldsymbol{\mu}d\boldsymbol{\tau}d\boldsymbol{\theta} \leq \tilde{M} \int_{\Omega_1} \pi_0(\boldsymbol{\theta}\,|\,\mathbf{x})d\boldsymbol{\theta} \int_{\Omega_2} d\boldsymbol{\mu} \int_{\Omega_3} d\boldsymbol{\tau} < \infty$ as required.

APPENDIX B

The prior density functions for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are given below. Clearly, they are density functions truncated on the parameter space of the posterior density function.

$$\pi(\boldsymbol{\alpha}) = \prod_{j=0}^{k} \frac{1}{\sqrt{2\pi}\ \sigma_j} e^{-\alpha_j^2/2\sigma_j^2} \left\{ \int_{-M}^{M} \frac{1}{\sqrt{2\pi}\ \sigma_j} e^{-\alpha_j^2/2\sigma_j^2} d\alpha_j \right\}^{-1},$$

$$\pi(\boldsymbol{\beta}) = \prod_{j=0}^{k} \frac{1}{\sqrt{2\pi}\ \beta_j s_j} e^{-\ln^2(\beta_j)/2s_j^2} \left\{ \int_{\epsilon}^{M} \frac{1}{\sqrt{2\pi}\ \beta_j s_j} e^{-\ln^2(\beta_j)/2s_j^2} d\beta_j \right\}^{-1},$$

$$\pi(\boldsymbol{\gamma}) = \prod_{v=1}^{2} \frac{1}{\sqrt{2\pi}\lambda_v(-\gamma_v)} e^{-(\ln(-\gamma_v))^2/2\lambda_v^2} \left\{ \int_{\epsilon}^{M} \frac{1}{\sqrt{2\pi}\lambda_v(-\gamma_v)} e^{-(\ln(-\gamma_v))^2/2\lambda_v^2} d(-\gamma_v) \right\}^{-1},$$

where σ_j, s_j and λ_v are the scale parameters of the respective prior density functions.

APPENDIX C

Our MCMC method is given below. Let $\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ represent the current and $\boldsymbol{\mu}'$, $\boldsymbol{\tau}'$, $\boldsymbol{\alpha}'$, $\boldsymbol{\beta}'$, $\boldsymbol{\gamma}'$ represent the proposed parameter values. Then our MCMC method consists of the following steps.

Step 1 Propose α_j' by simulating $\alpha_j' \sim \left(\tilde{\sigma}_j\sqrt{2\pi}\right)^{-1} e^{-(\alpha_j'-\alpha_j)^2/2\tilde{\sigma}_j^2}$ such that $\alpha_j' \in [-M, M]$, where $j = 0, \ldots, k$.

Step 2 Propose β_j' by simulating $\beta_j' \sim \left(\tilde{s}_j\beta_j'\sqrt{2\pi}\right)^{-1} e^{-(\ln\beta_j'-\ln\beta_j)^2/2\tilde{s}_j^2}$ such that $\beta_j' \in [\epsilon, M]$, where $j = 0, \ldots, k$.

Step 3 Propose γ_v' by simulating $-\gamma_v' \sim \left\{\tilde{\lambda}_v(-\gamma_v')\sqrt{2\pi}\right\}^{-1} e^{-\{\ln(-\gamma_v')-\ln(-\gamma_v)\}^2/2\tilde{\lambda}_v^2}$ such that $\gamma_v' \in [-M, -\epsilon]$, where $v = 1, 2$.

Step 4 Propose $\tau_i' = 0.5$ and $\mu_i' = e^{\eta_i'}/(1+e^{\eta_i'})$, where $\eta_i' = (\alpha_0' + \alpha_1'x_{1i} + \cdots + \alpha_k'x_{ki}) + \left(\beta_0' + \beta_1'x_{1i}^2 + \cdots + \beta_k'x_{ki}^2\right)\left\{\frac{0.5^{\gamma_1'}-1}{\gamma_1'} - \frac{0.5^{\gamma_2'}-1}{\gamma_2'}\right\}$.

Step 5 Accept the proposed values with probability $\min\{AB, 1\}$, where A and B are given below.

Step 6 If the proposed values are accepted, let $(\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = (\boldsymbol{\mu}', \boldsymbol{\tau}', \boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}')$. Otherwise, discard $(\boldsymbol{\mu}', \boldsymbol{\tau}', \boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}')$. Go to Step 1.

The acceptance probability of the MCMC method is given by $\min\{AB, 1\}$, where

$$
\begin{aligned}
A &= \frac{\pi(\boldsymbol{\mu}', \boldsymbol{\tau}', \boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}' | \mathbf{x}, \mathbf{y})}{\pi(\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma} | \mathbf{x}, \mathbf{y})} \\
&= \prod_{i=1}^{n} \frac{(\mu_i')^{y_i}(1-(\mu_i'))^{1-y_i}}{\mu_i^{y_i}(1-\mu_i)^{1-y_i}} \ \frac{\pi(\boldsymbol{\mu}' | \boldsymbol{\tau}', \boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}', \mathbf{x})}{\pi(\boldsymbol{\mu} | \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{x})} \ \frac{\pi(\boldsymbol{\tau}' | \boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}', \mathbf{x})}{\pi(\boldsymbol{\tau} |, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{x})} \ \frac{\pi(\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}' | \mathbf{x})}{\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma} | \mathbf{x})} \\
&= \prod_{i=1}^{n} \frac{(\mu_i')^{y_i}(1-(\mu_i'))^{1-y_i}}{\mu_i^{y_i}(1-\mu_i)^{1-y_i}} \ \frac{\pi(\boldsymbol{\mu}' | \boldsymbol{\tau}', \boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}', \mathbf{x})}{\pi(\boldsymbol{\mu} | \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{x})} \ \frac{\pi(\boldsymbol{\alpha}')\pi(\boldsymbol{\beta}')\pi(\boldsymbol{\gamma}')}{\pi(\boldsymbol{\alpha})\pi(\boldsymbol{\beta})\pi(\boldsymbol{\gamma})}
\end{aligned}
$$

$$
\begin{aligned}
B &= \frac{q\{(\boldsymbol{\mu}', \boldsymbol{\tau}', \boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}' | \mathbf{x}) \to (\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma} | \mathbf{x})\}}{q\{(\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma} | \mathbf{x}) \to (\boldsymbol{\mu}', \boldsymbol{\tau}', \boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}' | \mathbf{x})\}} = \frac{\pi(\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma} | \boldsymbol{\mu}', \boldsymbol{\tau}', \boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}', \mathbf{x})}{\pi(\boldsymbol{\mu}', \boldsymbol{\tau}', \boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}' | \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{x})} \\
&= \frac{\begin{array}{c}\pi(\boldsymbol{\mu} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\tau}, \boldsymbol{\tau}', \boldsymbol{\mu}', \boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}', \mathbf{x})\pi(\boldsymbol{\alpha} | \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\mu}', \boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}', \mathbf{x}) \\ \pi(\boldsymbol{\beta} | \boldsymbol{\gamma}, \boldsymbol{\mu}', \boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}', \mathbf{x})\pi(\boldsymbol{\gamma} | \boldsymbol{\mu}', \boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}', \mathbf{x})\end{array}}{\begin{array}{c}\pi(\boldsymbol{\mu}' | \boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}', \boldsymbol{\tau}, \boldsymbol{\tau}', \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{x})\pi(\boldsymbol{\alpha}' | \boldsymbol{\beta}', \boldsymbol{\gamma}', \boldsymbol{\mu}, \\ \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{x})\pi(\boldsymbol{\beta}' | \boldsymbol{\gamma}', \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{x})\pi(\boldsymbol{\gamma}' | \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{x})\end{array}} \\
&= \frac{\pi(\boldsymbol{\mu} | \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{x})}{\pi(\boldsymbol{\mu}' | \boldsymbol{\tau}', \boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}', \mathbf{x})} \ \frac{\pi(\boldsymbol{\alpha} | \boldsymbol{\alpha}')}{\pi(\boldsymbol{\alpha}' | \boldsymbol{\alpha})} \ \frac{\pi(\boldsymbol{\beta} | \boldsymbol{\beta}')}{\pi(\boldsymbol{\beta}' | \boldsymbol{\beta})} \ \frac{\pi(\boldsymbol{\gamma} | \boldsymbol{\gamma}')}{\pi(\boldsymbol{\gamma}' | \boldsymbol{\gamma})}.
\end{aligned}
$$

Hence,

$$
AB = \prod_{i=1}^{n} \frac{(\mu_i')^{y_i}(1-(\mu_i'))^{1-y_i}}{\mu_i^{y_i}(1-\mu_i)^{1-y_i}} \ \frac{\pi(\boldsymbol{\alpha}')\pi(\boldsymbol{\beta}')\pi(\boldsymbol{\gamma}')}{\pi(\boldsymbol{\alpha})\pi(\boldsymbol{\beta})\pi(\boldsymbol{\gamma})} \ \frac{\pi(\boldsymbol{\alpha} | \boldsymbol{\alpha}')}{\pi(\boldsymbol{\alpha}' | \boldsymbol{\alpha})} \ \frac{\pi(\boldsymbol{\beta} | \boldsymbol{\beta}')}{\pi(\boldsymbol{\beta}' | \boldsymbol{\beta})} \ \frac{\pi(\boldsymbol{\gamma} | \boldsymbol{\gamma}')}{\pi(\boldsymbol{\gamma}' | \boldsymbol{\gamma})},
$$

where

$$
\frac{\pi(\boldsymbol{\alpha}')\pi(\boldsymbol{\beta}')\pi(\boldsymbol{\gamma}')}{\pi(\boldsymbol{\alpha})\pi(\boldsymbol{\beta})\pi(\boldsymbol{\gamma})} = \prod_{j=0}^{k} e^{-(\alpha_j'^2 - \alpha_j^2)/2\sigma_j^2} \ \frac{\beta_j}{\beta_j'} e^{-\left\{\ln^2(\beta_j') - \ln^2(\beta_j)\right\}/2s_j^2} \prod_{v=1}^{2} \frac{\gamma_v}{\gamma_v'} e^{-\left\{\ln^2(-\gamma_v') - \ln^2(-\gamma_v)\right\}/2\lambda_v^2},
$$

$$
\frac{\pi(\boldsymbol{\alpha} | \boldsymbol{\alpha}')}{\pi(\boldsymbol{\alpha}' | \boldsymbol{\alpha})} = 1, \quad \frac{\pi(\boldsymbol{\beta} | \boldsymbol{\beta}')}{\pi(\boldsymbol{\beta}' | \boldsymbol{\beta})} = \prod_{j=0}^{k} \frac{\beta_j'}{\beta_j} \frac{\Phi\left(\frac{\ln M - \ln \beta_j}{\tilde{s}_j}\right) - \Phi\left(\frac{\ln \epsilon - \ln \beta_j}{\tilde{s}_j}\right)}{\Phi\left(\frac{\ln M - \ln \beta_j'}{\tilde{s}_j}\right) - \Phi\left(\frac{\ln \epsilon - \ln \beta_j'}{\tilde{s}_j}\right)}
$$

$$
\frac{\pi(\boldsymbol{\gamma} | \boldsymbol{\gamma}')}{\pi(\boldsymbol{\gamma}' | \boldsymbol{\gamma})} = \prod_{v=1}^{2} \frac{\gamma_v'}{\gamma_v} \frac{\Phi\left(\frac{\ln M - \ln(-\gamma_v)}{\tilde{\lambda}_v}\right) - \Phi\left(\frac{\ln \epsilon - \ln(-\gamma_v)}{\tilde{\lambda}_v}\right)}{\Phi\left(\frac{\ln M - \ln(-\gamma_v')}{\tilde{\lambda}_v}\right) - \Phi\left(\frac{\ln \epsilon - \ln(-\gamma_v')}{\tilde{\lambda}_v}\right)}.
$$

Therefore,

$$
\begin{aligned}
AB = &\prod_{i=1}^{n} \frac{(\mu_i')^{y_i}(1-(\mu_i'))^{1-y_i}}{\mu_i^{y_i}(1-\mu_i)^{1-y_i}} \prod_{j=0}^{k} e^{-(\alpha_j'^2 - \alpha_j^2)/2\sigma_j^2} \ e^{-\left\{\ln^2(\beta_j') - \ln^2(\beta_j)\right\}/2s_j^2} \\
&\times \prod_{v=1}^{2} e^{-\left\{\ln^2(-\gamma_v') - \ln^2(-\gamma_v)\right\}/2\lambda_v^2} \prod_{j=0}^{k} \frac{\Phi\left(\frac{\ln M - \ln \beta_j}{\tilde{s}_j}\right) - \Phi\left(\frac{\ln \epsilon - \ln \beta_j}{\tilde{s}_j}\right)}{\Phi\left(\frac{\ln M - \ln \beta_j'}{\tilde{s}_j}\right) - \Phi\left(\frac{\ln \epsilon - \ln \beta_j'}{\tilde{s}_j}\right)} \\
&\times \prod_{v=1}^{2} \frac{\Phi\left(\frac{\ln M - \ln(-\gamma_v)}{\tilde{\lambda}_v}\right) - \Phi\left(\frac{\ln \epsilon - \ln(-\gamma_v)}{\tilde{\lambda}_v}\right)}{\Phi\left(\frac{\ln M - \ln(-\gamma_v')}{\tilde{\lambda}_v}\right) - \Phi\left(\frac{\ln \epsilon - \ln(-\gamma_v')}{\tilde{\lambda}_v}\right)}.
\end{aligned}
$$

APPENDIX D

Another simulation study

In this simulation study, we show that the proposed estimation method performs well, and the convergence of the method does not depend on the strength of the prior information on the parameters.

For $i = 1, \ldots, 500$, we simulated x_i uniformly on [0, 5] and η_i from

$$Q_{\eta_i}(\tau|x_i) = (-0.4 + 0.2x_i) + (2 + 0.3x_i^2)\left\{\frac{\tau^{-0.35} - 1}{-0.35} - \frac{(1-\tau)^{-0.5} - 1}{-0.5}\right\}, \qquad \text{(D1)}$$

which can be achieved by simulating τ_i uniformly on (0, 1) and calculating $\eta_i = Q_{\eta_i}(\tau_i|x_i)$ using (D1). Then μ_i values can be calculated by $\mu_i = e^{\eta_i}/(1 + e^{\eta_i})$. Finally, we let $y_i = 1$ with probability μ_i and $y_i = 0$ otherwise. By repeating these steps 200 times, we obtained 200 independent data sets.

Note that in reality both η_i and μ_i are not observable, but here we will use them to check the performance of our estimation method. Specifically, we will (a) compare the distribution of the estimated standardised residuals $\hat{r}_i = \{\eta_i - (\hat{\alpha}_0 + \hat{\alpha}_1 x_i)\}/\{\hat{\beta}_0 + \hat{\beta}_1 x_i^2\}$ ($i = 1, \ldots, 500$) with the true one

$$Q(\tau, \boldsymbol{\gamma}) = (\tau^{-0.35} - 1)/(-0.35) - \{(1-\tau)^{-0.5} - 1\}/(-0.5); \qquad \text{(D2)}$$

(b) compare the estimated distribution of μ with the true one defined by (7); and (c) check the coverage probabilities of the estimated quantile curves of η and μ respectively. If the estimation method performs well, we should expect a good agreement between the estimated results and true results.

It is worth mentioning that the strength of the prior information about the parameters of the posterior density function is measured by the values of σ_j, s_ℓ and λ_v (see Appendix B). For example, a large (or small) value of σ_j suggests that the prior information on α_j is weak (or strong). To check the effect of the prior information on the estimation, we let $\sigma_j = s_\ell = \lambda_v = \xi$, where $\xi = 1, 2, 3$. So, the variance of α_j is given by 1, 4 and 9 for $\xi = 1, 2$ and 3 respectively, while the variances of β_ℓ and γ_v are the same, given by 4.671, 2926 and 6.565×10^7 corresponding to $\xi = 1, 2$ and 3 respectively. Clearly, when $\xi = 2$ the strength of the prior information has already become very weak.

Now, for each value of ξ, a Markov Chain was run for 5×10^5 steps. Testing runs suggest that a burn-in period of the first 2×10^4 steps is enough. The posterior samples were then collected after the burn-in period. The first two rows of Figure D1 show the time series plots of the posterior samples obtained by using the first simulated data set and $\xi = 2$, which suggests that the convergence of the Markov Chain has been achieved. The last two rows of Figure D1 show the probability density function plots of the posterior marginal distributions of the parameters, where the vertical lines correspond to the true parameter values, and the darker continuous, grey continuous and dotted curves correspond to priors with $\xi = 1, 2$ and 3 respectively. It is seen that in all cases the true parameter values are well within the range of the posterior marginal distributions. To save space, we will present our results corresponding to $\xi = 2$ in the paper.

For each data set, we recorded the Bayesian estimate together with an associated 95% credible interval, leading to 200 credible intervals for each parameter. Table D1 shows the

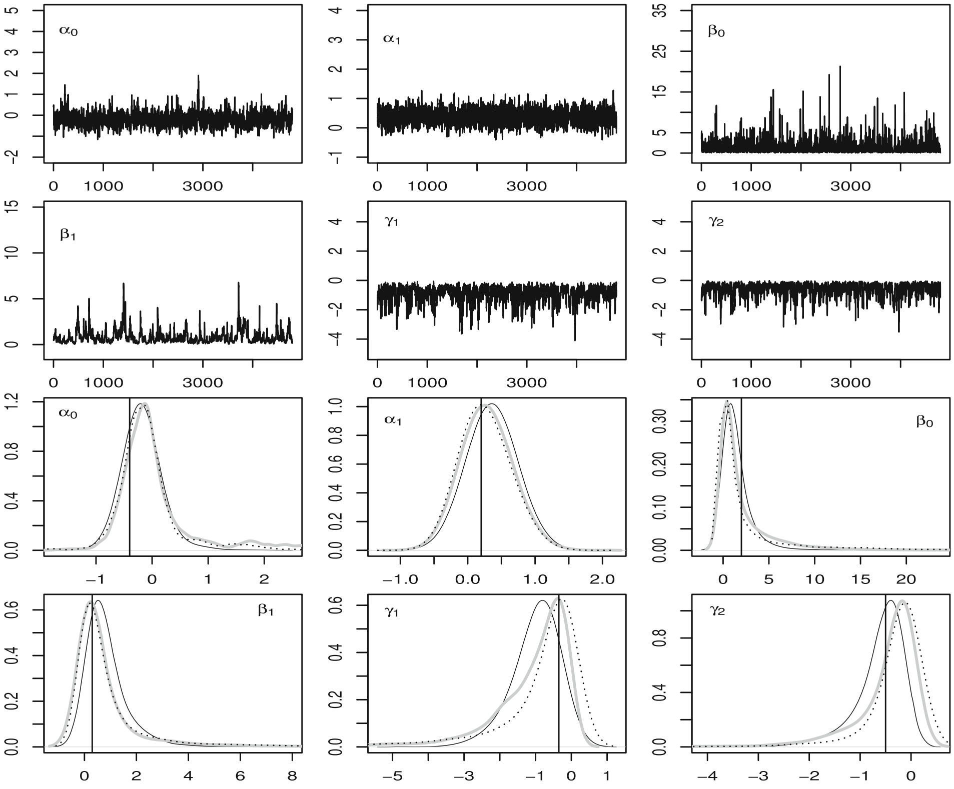

Figure D1. Top two rows: Time series plots of the posterior samples of the model parameters for $\xi = 2$. Last two rows: Plots of the posterior marginal density functions of the model parameters, where the darker continuous, grey continuous and dotted curves correspond to $\xi = 1, 2$ and 3 respectively, and the vertical lines correspond to the true model parameters.

Table D1. Estimation results.

Parameter	α_0	α_1	β_0	β_1	γ_1	γ_2
True value	−0.40	0.20	2.00	0.30	−0.35	−0.50
Lower bound	−1.120	−0.287	0.019	0.012	−2.705	−2.461
Upper bound	1.237	0.494	22.268	5.086	−0.019	−0.014

true parameter values and the average value of the lower (upper) bounds of these credible intervals. It is seen that all the true parameter values are well within the respective lower and upper bounds, suggesting a good performance of the method.

We now compare the distribution of $\hat{r}_i$ with (D2). A good performance of the method is expected if the two distributions are not significantly different. So we estimated a probability density function by using $\hat{r}_i$ $(i = 1, \ldots, 500)$ for each data set, which is then compared with (D2) by using the Kolmogorov−Smirnov test. We found that the average p-value of the 200 tests is 0.0642 with a 95% confidence interval [0.0454, 0.0831]. Hence the Kolmogorov−Smirnov test shows that, on average, the distribution of $\hat{r}_i$ is not significantly different from the true distribution defined by (D2) at a 1% level of significance.

Similarly, we compared the estimated and true distributions of μ. We found that the average p-value of the Kolmogorov–Smirnov tests over 200 simulations is 0.0433 with a 95% confidence interval ranging from 0.0313 to 0.0554. Hence, these two distributions are also not significantly different at a 1% level of significance, providing further evidence about the good performance of the method.

Let us now consider the estimated conditional quantiles of η and μ. For $\tau \in \{0.025, 0.05,\ 0.25, 0.5, 0.75, 0.95, 0.975\}$, let $n^{\tau}_{\eta\ell}$ (or $n^{\tau}_{\eta\ell}/500$) and $n^{\tau}_{\mu\ell}$ (or $n^{\tau}_{\mu\ell}/500$) be the number (or proportion) of observed η_i and μ_i in the ℓth simulated data set that are less than $\hat{Q}_{\eta_i}(\tau|\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \mathbf{x}_i)$ and $\hat{Q}_{\mu_i}(\tau|\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \mathbf{x}_i)$ respectively, where $\ell = 1, \ldots, 200$. For each ℓ, we further calculated mean squared error (MSE) between $n^{\tau}_{\eta\ell}/500$ and τ and between $n^{\tau}_{\mu\ell}/500$ and τ respectively. A good performance of the method is expected if these MSE values are all small for $\ell = 1, \ldots, 200$. The average MSE value for η is 0.0065 with a 95% confidence interval $[0.006, 0.0069]$ and that for μ is 0.0051 with a 95% confidence interval $[0.0049, 0.0054]$. Clearly, both of these MSEs are very small. In summary, the performance of the estimation method is satisfactory in this study.

References

ACHEN, C.H. (2002). Toward a new political methodology: Microfoundations and ART. *Annual Review of Political Science*, **5**, 423–450.

ALBERS, S. (2012). Optimizable and implementable aggregate response modeling for marketing decision support. *International Journal of Research in Marketing*, **29**, 111–122.

ALVAREZ, R.M. & BREHM, J. (1995). Ambivalence towards abortion policy: Development of a heteroskedastic probit model of competing values. *American Journal of Political Science*, **39**, 1055–1082.

BENOIT, D.F. & VAN DEN POEL, D. (2012). Binary quantile regression: A Bayesian approach based on the asymmetric Laplace distribution. *Journal of Applied Econometrics*, **27**, 1174–1188.

BHATTACHARYA, S. (1999). Direct marketing performance modeling using genetic algorithms. *INFORMS Journal on Computing*, **11**, 248–257.

BRANCO, P., TORGO, L. & RIBEIRO, R.P. (2016). A survey of predictive modelling on imbalanced domains. *ACM Computing Surveys*, **49**, 1–50.

BREIMAN, L., FRIEDMAN, J., OLSHEN, R. & STONE, C. (1984). *Classification and Regression Trees*. New York: Chapman and Hall/CRC.

BREIMAN, L. (2001). Random forests. *Machine Learning*, **45**, 5–32.

BROOKS, S.P. (1998). Markov chain Monte Carlo method and its application. *The Statistician*, **47**, 69–100.

BRUNO, H.A., CEBOLLADA, J. & CHINTAGUNTA, P.K. (2018). Targeting Mr. or Mrs. Smith: Modeling and leveraging intrahousehold heterogeneity in brand choice behavior. *Marketing Science*, **37**, 631–648.

BULT, J.R. & WANSBEEK, T.J. (1995). Optimal selection for direct mail. *Marketing Science*, **14**, 378–394.

BULT, J.R. (1993). Semiparametric versus parametric classification models: An application to direct Marketing. *Journal of Marketing Research*, **30**, 380–390.

CORTES, C. & VAPNIK, V. (1995). Support-vector networks. *Machine Learning*, **20**, 273–297.

COVER, T.M. & HART, P.E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, **13**(1), 21–27.

CUI, G., WONG, L. & WAN, X. (2012). Cost-sensitive learning via priority sampling to improve the return on marketing and CRM investment. *Journal of Management Information Systems*, **29**, 341–374.

FAWCETT, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, **27**, 861–874.

FOURNIER, B., RUPIN, N., BIGERELLE, M., NAJJAR, D., IOST, A. & WILCOX, R. (2007). Estimating the parameters of a generalized lambda distribution. *Computational Statistics & Data Analysis*, **51**, 2813–2835.

GEYER, C.J. (2011). Introduction to Markov Chain Monte Carlo. In *Handbook of Markov Chain Monte Carlo*, eds., Brooks, S., Gelman, A., Jones, G.K., & Meng, X.L., pp. 3–48. FL: CRC Press.

GILCHRIST, W.G. (2000). *Statistical Modelling with Quantile Functions*. New York: Chapman & Hall/CRC.

HASHEM, H., VINCIOTTI, V., ALHAMZAWI, R. & YU, K. (2016). Quantile regression with group lasso for classification. *Advances in Data Analysis and Classification*, **10**, 375–390.

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J.H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn. New York: Springer.

HOROWITZ, J.L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica*, **60**, 505–531.

KAPPE, E., BLANK, A.S. & DESARBO, W.S. (2018). A random coefficients mixture hidden Markov model for marketing research. *International Journal of Research in Marketing*, **35**, 415–431.

KOENKER, R. (2005). *Quantiles Regression*. Cambridge: Cambridge University Press.

KOHAVI, R. & PROVOST, F. (1998). Glossary of Terms. *Machine Learning*, **30**, 271–274.

KORDAS, G. (2006). Smoothed binary regression quantiles. *Journal of Applied Econometrics*, **21**, 387–407.

LI, Y. & ANSARI, A. (2014). A Bayesian semiparametric approach for endogeneity and heterogeneity in choice models. *Management Science*, **60**, 1161–1179.

LING, C.X. & LI, C. (1998). Data mining for direct marketing: Problems and solutions. In *Proceedings of the 4th KDD Conference*, pp. 73–79. New York: *AAAI Press*.

MANSKI, C.F. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics*, **3**, 205–228.

MANSKI, C.F. (1985). Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator. *Journal of Econometrics*, **27**, 313–333.

MCCULLAGH, P. & NELDER, J.A. (1989). *Generalized Linear Models*, 2nd edn. Boca Raton: Chapman & Hall/CRC.

MEINSHAUSEN, N. (2006). Quantile Regression Forests. *Journal of Machine Learning Research*, **7**, 983–999.

MORO, S., LAUREANO, R. & CORTEZ, P. (2011). Using data mining for bank direct marketing: an application of the CRISP-DM methodology. In *Proceedings of the European Simulation and Modelling Conference–ESM'2011*, eds., Novais, P. et al., pp. 117–121. Portugal: EUROSIS.

ROSSI, P.E., MCCULLOCH, R.E. & ALLENBY, G.M. (1996). The value of purchase history data in target marketing. *Marketing Science*, **15**, 321–340.

TRAIN, K. (2002). *Discrete Choice Methods with Simulation*. Berkeley: University of California.

VUK, M. & CURK, T. (2006). roc curve, lift chart and calibration plot. *Metodološki zvezki*, **3**, 89–108.

WANG, M., GOU, Q., WU, C. & LIANG, L. (2013). An aggregate advertising response model based on consumer population dynamics. *International Journal of Applied Management Science*, **5**, 22–38.

YATCHEW, A. & GRILICHES, Z. (1985). Specification error in probit models. *Review of Economics and Statistics*, **18**, 134–139.

ZAHAVI, J. & LEVIN, N. (1997). Applying neural computing to target marketing. *Journal of Direct Marketing*, **11**, 76–93.