Hindawi
Discrete Dynamics in Nature and Society
Volume 2022, Article ID 3112987, 13 pages
https://doi.org/10.1155/2022/3112987

Research Article

Promoting Variable Effect Consistency in Mixture Cure Model for Credit Scoring

Chenlu Zheng,[1,2] **Jianping Zhu**,[1,2] **Xinyan Fan**,[3] **Song Chen**,[4] **and Zhiyuan Zhang**[5]

[1]*School of Management, Xiamen University, Xiamen 361005, China*
[2]*Data Mining Research Center, Xiamen University, Xiamen 361005, China*
[3]*School of Statistics, Renmin University of China, Beijing 100872, China*
[4]*Taizhou University, Microfinance College, Taizhou 318000, China*
[5]*Science and Technology Development Department, Xiamen International Bank, Xiamen 361001, China*

Correspondence should be addressed to Jianping Zhu; xmzhujianping@163.com and Song Chen; chensong1978@tzc.edu.cn

Received 3 October 2021; Revised 10 December 2021; Accepted 31 December 2021; Published 21 February 2022

Academic Editor: Dehua Shen

Mixture cure models are widely adopted in credit scoring. Mixture cure models consist of two parts: an incident part which predicts the probability of default and a latency part which predicts when they are likely to default. The two model parts describe two quite relevant credit aspects. So, it is reasonable to expect that the two sets of the coefficients are somewhat related. Moreover, in practical cases, it is difficult to interpret the results when the two sets of the coefficients of the same variables have conflicting signs. Most existing works either ignore the interconnections of the two sets of coefficients or impose a strict constraint between them. We proposed a mixture cure model considering the variable effect consistency using a sign-based penalty. It is a more flexible model that allows the two sets of coefficients to be in different distributions and magnitudes. To accommodate high-dimensional credit data, a group lasso penalty is also imposed for variable selection. Simulation shows that the proposed method has competitive performance compared with alternative methods in terms of estimation and prediction. Furthermore, the empirical study illustrates that the proposed method outperforms the alternative method and can improve the interpretability of the results.

1. Introduction

Credit scoring is an effective and crucial approach for evaluating credit risk [1, 2]. A slight improvement in the prediction precision of credit scoring models can bring considerable benefits. Therefore, credit scoring has attracted increasing attention of scholars and practitioners. Many studies treat it as a classification problem to distinguish noncreditworthy customers from creditworthy ones [2, 3]. These studies focus on classification techniques including logistic regression, support vector machine, neural network, and random forest [4, 5]. For example, Zhou et al. [6] proposed a logistic regression method with clustering analysis for credit risk evaluation. Zhang et al. [7] proposed a cost-sensitive logistic regression model to assess the credit risk. Considering the high cost and time consumption of credit scoring, a credit granting process using three-way decisions is proposed to make efficient credit decisions [8]. Since the exposure to risk and the losses caused by default are strongly related to the time when they default [9], credit risk prediction overtime is of great significance for timely risk management.

Survival analysis, with its ability of predicting the probability of default over time, has been first applied in credit scoring in 1992 [10]. It can be more informative than the binary classification model. Subsequently, various survival analysis models are proposed to predict credit risk over time. For instance, Cox proportional hazards (PH) model is adopted to predict early repayment and time to default in personal loans and investigated the effect of different variables on time to default [11, 12]. In addition, macroeconomic factors and time-varying data are also incorporated in

survival analysis to improve the performance of prediction in credit scoring [13, 14]. And the models are further extended by a survival gradient boosting decision tree approach to enhance the prediction performance [15].

However, standard survival analysis assumes that the loan term is long enough and every customer will eventually default. In practice, a substantial proportion of customers will not default during the entire loan term. Since mixture cure models applied in medicine assume that some patients have been cured and will not die during the follow-up period, it is more appropriate in the credit market and was first introduced to credit scoring by Tong et al. [16].

Recently, the mixture cure model, an extension of the standard survival analysis, is widely adopted in credit scoring for its ability of predicting not only whether customers will default but also when they are likely to default. Results showed that the mixture cure model is more suitable for credit data compared with standard survival analysis models and the mixture cure model incorporating penalized spline has better performance in prediction [17]. Mixture cure models have been further developed by identifying different risk patterns of customers, considering the influence of competitive risk, and the relationship between the default times and the variables [18–20].

Mixture cure models consist of two parts: an incident part which predicts the probability of default and a latency part which predicts when they are likely to default. In the two model parts, the two sets of the coefficients indicate the two sets of the variable effects on the credit risk. The two model parts describe two quite relevant credit aspects. Nevertheless, most of the existing studies ignore the relations between the two sets of coefficients in two model parts. These works generally assume that there are no direct constraints between the two sets of coefficients, which may get conflicting results of variable effects. For example, Dirick et al. [21] propose a mixture cure model incorporating macroeconomic factors to predict credit risk. The results show that the customers' annual income has the opposite effect on whether and when to default. In other words, according to the results, customers with lower annual income have a lower probability of default, but they are more likely to default earlier. It is difficult to interpret the conflicting results and apply them in practice.

In fact, the two model parts describe two quite relevant credit aspects, namely, the probability of default and survival (nondefault) time. Customers with high default probability are more likely to default earlier. So, it is reasonable to expect that the two sets of the coefficients are somewhat related. Theoretical derivations [22] and empirical analysis [23] also suggest that relaxing the independence of two sets of coefficients can improve the model performance. The assumption has been relaxed by establishing a joint distribution of the defaulting predictor and the logarithm of the hazard rate in [23]. Note that the two model parts still describe two different aspects of default. The assumption of a joint distribution may be too strict. So, we consider a more flexible model that allows the two sets of coefficients to be in different distributions and magnitudes. Sign consistency penalty is proposed to promote the similarity in sign to get more interpretable results by Zhang et al. [24]. In this paper, we propose a variable effect consistency mixture cure model with a sign-based penalty. The proposed method can promote the similarity in the signs of variable effect in the two model parts to improve interpretability. To accommodate high-dimensional credit data, a group lasso penalty is also imposed for variable selection [25].

The contributions of this paper are as follows. First, we propose a variable effect consistency mixture cure model. The proposed method can lead to more interpretable results by promoting the similarity in the signs of coefficients in the two parts of the mixture cure model. Second, a group lasso penalty is imposed to select important variable subgroups and accommodate the high-dimensional data. Third, simulation and empirical analysis of credit data illustrate that the proposed method can improve the prediction accuracy as well as interpretability, which has important practical significance for applying the prediction results to the credit business.

The remainder of the paper is organized as follows. Section 2 introduces the variable effect consistency mixture cure model. Computational algorithm is presented in Section 3. Simulation is carried out in Section 4. Empirical study is presented in Section 5. Finally, conclusions are discussed in Section 6.

2. Methods

In this paper, we consider credit data with n customers and p variables. Denote U as the time to default and C as the time of censoring. Let Y be the unobservable binary variable with $Y = 0$ indicating that the customer is cured and will not default ($U = \infty$), whereas $Y = 1$ indicates an uncured customer and will eventually default.

Denote $\delta_i = \mathrm{I}(U_i < C_i)$ as the censoring indicator of customer i, where U_i and C_i are the time to default and censoring time of customer i, respectively. $\delta_i = 0$ if censored and $\delta_i = 1$ otherwise. Note that there are three possible credit states of customers. (a) $\delta_i = 0$ and $Y_i = 0$: censored, cured customers who will not default; (b) $\delta_i = 0$ and $Y_i = 1$: censored, uncured customers who will eventually default and have not been observed to default in censoring time C_i; (c) $\delta_i = 1$ and $Y_i = 1$: uncensored, uncured customers who have been observed to default.

Denote $t_i = \min(U_i, C_i)$. Note that, in many practical cases, variables can be naturally grouped. For instance, many categorical variables may have several levels and can be represented by subgroups of dummy variables [26]. The additive model with polynomial or nonparametric components can be expressed as groups of basis functions [27]. In addition, grouping structure can also be introduced by taking advantage of prior knowledge. For example, genes belonging to the same biological pathway can also be considered a group [28]. Let $\mathbf{x} = (\mathbf{x}_1^{\mathrm{T}}, \ldots, \mathbf{x}_J^{\mathrm{T}})^{\mathrm{T}}$ be the variable vector with J subgroups. $\mathbf{x}_j = (x_{j1}, \ldots, x_{jp_j})^{\mathrm{T}}$ is the j-th subgroup of variable vector, and $\sum_{j=1}^{J} p_j = p$. The observable data are $(t_i, \delta_i, \mathbf{x})$, $i = 1, \ldots, n$.

The incident part of the mixture cure model describes the probability of default, for which we adopt a logistic

regression model. Let π_i be the probability of cured (non-default) customer i.

$$\pi_i = \Pr(Y_i = 0|\mathbf{x}) = \frac{1}{1+\exp(\alpha_0 + \mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha})}, \tag{1}$$

where α_0 is the intercept, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^{\mathrm{T}}, \ldots, \boldsymbol{\alpha}_J^{\mathrm{T}})^{\mathrm{T}}$ is the vector of unknown regression coefficients, and the j-th subgroup of the coefficient vector is $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jp_j})^{\mathrm{T}}$.

In the latency part, for uncured customers, we adopt an exponential model for survival. Note that the exponential model has been commonly adopted in mixture cure models [29, 30]. It is easy to capture the relations between the probability and the time to default for it includes only one parameter [23]. The survival function is

$$\begin{aligned} S_i(t|Y_i = 1, \mathbf{x}) &= \exp(-\lambda_i t), \\ \lambda_i &= \exp(\beta_o + \mathbf{x}^T\boldsymbol{\beta}), \end{aligned} \tag{2}$$

where λ_i is the hazard function of customer i, β_0 is the intercept, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_J^{\mathrm{T}})^T$ is the vector of unknown regression coefficients, and the j-th subgroup of the coefficient vector is $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp_j})^{\mathrm{T}}$. Survival function $S_i(t|Y_i = 1, \mathbf{x}) = P(U_i > t|Y_i = 1, \mathbf{x})$ indicates the survival probability of uncured customers in time t, that is, the probability of default in time t given the customer will default.

The mixture cure model can be given by

$$S_i(t|\mathbf{x}) = \pi_i + (1-\pi_i)S_i(t|Y_i = 1, \mathbf{x}), \tag{3}$$

where $S_i(t|\mathbf{x})$ is the survival probability of customer i in time t.

For observable data $(t_i, \delta_i, \mathbf{x})$, the objective function can be written as follows.

$$Q(\alpha_0, \boldsymbol{\alpha}, \beta_0, \boldsymbol{\beta}) = -L(\alpha_0, \boldsymbol{\alpha}, \beta_0, \boldsymbol{\beta}) + P(\boldsymbol{\alpha}, \boldsymbol{\beta}), \tag{4}$$

where $L(\alpha_0, \alpha, \beta_0, \beta)$ is the log-likelihood function, which is

$$L(\alpha_0, \boldsymbol{\alpha}, \beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} \begin{pmatrix} \delta_i(\log(1-\pi_i) + \log(\lambda_i) - \lambda_i t_i) \\ +(1-\delta_i)\log((1-\pi_i)\exp(-\lambda_i t_i) + \pi_i) \end{pmatrix}, \tag{5}$$

with $\pi_i = 1/(1+\exp(\alpha_0 + \mathbf{x}^{\mathbf{T}}\boldsymbol{\alpha}))$, and $\lambda_i = \exp(\beta_o + \mathbf{x}^{\mathbf{T}}\boldsymbol{\beta})$. Here, $P(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is the penalty function, which is

$$\begin{aligned} P(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \mu_1 &\sum_{j=1}^{J} \sqrt{p_j}\left(\|\boldsymbol{\alpha}_j\| + \|\boldsymbol{\beta}_j\|\right) \\ &+ \frac{\mu_2}{2}\sum_{j=1}^{J}\sum_{k=1}^{p_j}\left(\operatorname{sign}(\alpha_{jk}) - \operatorname{sign}(\beta_{jk})\right)^2, \end{aligned} \tag{6}$$

where $\mu_1 > 0$ and $\mu_2 > 0$ are tuning parameters, $\|\cdot\|$ is the l_2 norm, and $\operatorname{sign}(\cdot)$ is the sign function. In many practical cases, grouping structure arises naturally. In addition, it is hard to interpret the results when coefficients corresponding to the same variables have conflicting signs. Therefore, we consider a flexible mixture cure model with sign consistency and group variable selection penalties. The first penalty is a group lasso penalty. It can conduct estimation and group variable selection by shrinking the coefficients of insignificant groups to 0. It considers grouping structures and has good prediction performance [26]. The second penalty is the sign consistency penalty. It promotes the sign consistency of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in the two parts of the model, which can lead to more interpretable results [31].

3. Computational Algorithm

In this section, the Expectation Coordinate Descent (ECD) algorithm is developed to optimize the objective function. In E-step, a latent unobserved Y_i is introduced to obtain a complete log-likelihood function. In CD-step, group coordinate descent is adopted to iteratively update a single parameter with the remaining parameters fixed at their most recent values. Sign function $\operatorname{sign}(\cdot)$ is difficult to optimize for its discontinuity and nondifferentiability. Therefore, referring to [24, 32], we propose the approximation as follows:

$$\left(\operatorname{sign}(\alpha_{jk}) - \operatorname{sign}(\beta_{jk})\right)^2 \approx \left(\frac{\alpha_{jk}}{|\alpha_{jk}| + \tau} - \frac{\beta_{jk}}{|\beta_{jk}| + \tau}\right)^2, \tag{7}$$

where τ is a small positive constant (more discussions below).

The ECD algorithm updates $(\alpha_0, \boldsymbol{\alpha}, \beta_0, \boldsymbol{\beta})$ in the m–th iteration as follows.

3.1. E-Step. Denote the observation of the latent Y_i as y_i and denote the complete data as $\{(t_i, \delta_i, y_i, \mathbf{x}), i = 1, \ldots, n\}$. The complete log-likelihood is

$$\begin{aligned} L^{[m]} &= \sum_{i=1}^{n}\left((1-y_i)\log(\pi_i^{[m]}) + y_i\log(1-\pi_i^{[m]})\right) \\ &\quad + \sum_{i=1}^{n}\left(\delta_i(\beta_0^{[m]} + \mathbf{x}^{\mathbf{T}}\boldsymbol{\beta}^{[m]}) - y_i t_i \exp(\beta_0^{[m]} + \mathbf{x}^{\mathbf{T}}\boldsymbol{\beta}^{[m]})\right) \\ &\coloneqq L_1^{[m]} + L_2^{[m]}, \end{aligned} \tag{8}$$

where

$$\pi_i^{[m]} = \frac{1}{1+\exp(\alpha_0^{[m]} + \mathbf{x}^{\mathbf{T}}\boldsymbol{\alpha}^{[m]})}. \tag{9}$$

The expectation of y_i is

$$E(y_i) = \begin{cases} 1 & \delta_i = 1 \\ \dfrac{(1-\pi_i^{[m]})\exp(-\lambda_i^{[m]} t_i)}{\pi_i^{[m]} + (1-\pi_i^{[m]})\exp(-\lambda_i^{[m]} t_i)} & \delta_i = 0 \end{cases}, \tag{10}$$

where

$$\lambda_i^{[m]} = \exp(\beta_0^{[m]} + \mathbf{x}^{\mathbf{T}}\boldsymbol{\beta}^{[m]}). \tag{11}$$

When customer i is observed to default ($\delta_i = 1$), the unobserved $y_i = 1$, whereas the expectation of y_i is related to the probability of cured and the uncured but censored customers.

In E-step, we take the expectation of $L^{[m]}$ with respect to $E(y_i)$ given the complete data $\{(t_i, \delta_i, y_i, \mathbf{x}), \quad i = 1, \cdots, n\}$.

$$E\left(L^{[m]}\right) = E\left(L_1^{[m]}\right) + E\left(L_2^{[m]}\right), \tag{12}$$

where

$$\begin{aligned} l_1^{[m]} &= E\left(L_1^{[m]}\right) = \sum_{i=1}^{n}\left((1 - E(y_i))\log\left(\pi_i^{[m]}\right) + E(y_i)\log\left(1 - \pi_i^{[m]}\right)\right), \\ l_2^{[m]} &= E\left(L_2^{[m]}\right) = \sum_{i=1}^{n}\left(\delta_i\left(\beta_0^{[m]} + \mathbf{x}^{\mathbf{T}}\boldsymbol{\beta}^{[m]}\right) - E(y_i)t_i \exp\left(\beta_0^{[m]} + \mathbf{x}^{\mathbf{T}}\boldsymbol{\beta}^{[m]}\right)\right). \end{aligned} \tag{13}$$

3.2. CD-Step. In CD-step, group coordinate descent is adopted to iteratively update $(\alpha_0, \boldsymbol{\alpha}, \beta_0, \boldsymbol{\beta})$. The intercept $\alpha_0^{[m+1]}$ is updated by

$$\alpha_0^{[m+1]} = \alpha_0^{[m]} - \left(\frac{\partial^2 l_1^{[m]}}{\partial\left(\alpha_0^{[m]}\right)^2}\right)^{-1} \frac{\partial l_1^{[m]}}{\partial \alpha_0^{[m]}}. \tag{14}$$

For $\boldsymbol{\alpha}_j^{[m+1]} \in R^{p_j}$, we adopted a fast unified algorithm, Groupwise Majorization Descent (GMD) proposed in [33] to solve the group lasso penalized objective function in (4). The upper bound of the objective function is as follows:

$$-\left(\boldsymbol{\alpha}_j^{[m+1]} - \boldsymbol{\alpha}_j^{[m]}\right)^{\mathrm{T}}\left(\frac{\alpha l_1^{[m]}}{\partial \boldsymbol{\alpha}_j^{[m]}} + \mu_2 V_{1j}^{[m]}\right) + \frac{M_{1j}^{[m]}}{2}\left(\boldsymbol{\alpha}_j^{[m+1]} - \boldsymbol{\alpha}_j^{[m]}\right)^{\mathrm{T}}\left(\boldsymbol{\alpha}_j^{[m+1]} - \boldsymbol{\alpha}_j^{[m]}\right)^{\mathrm{T}} + \mu_1\sqrt{p_j}\left\|\boldsymbol{\alpha}_j^{[m+1]}\right\|. \tag{15}$$

Here, $V_{1j}^{[m]}$ is p_j-length vector, and $M_{1j}^{[m]}$ is a constant as follows:

$$\begin{aligned} V_{1j}^{[m]} &= \left(\frac{1}{\left|\alpha_{jk}^{[m]}\right| + \tau}\left(\frac{\beta_{jk}^{[m]}}{\left|\beta_{jk}^{[m]}\right| + \tau} - \frac{\alpha_{jk}^{[m]}}{\left|\alpha_{jk}^{[m]}\right| + \tau}\right)\right)_{1 \le k \le p_j}, \\ M_{1j}^{[m]} &= \psi\left(-\left(\frac{\partial^2 l_1^{[m]}}{\partial \alpha_{jk_1}^{[m]} \partial \alpha_{jk2}^{[m]}}\right)_{1 \le k_1, k_2 \le p_j}\right) + \max_k\left(\mu_2\left(\frac{1}{\left|\alpha_{jk}^{[m]}\right| + \tau}\right)^2\right), \end{aligned} \tag{16}$$

where $\psi(\cdot)$ is the maximum eigenvalue function.

Similarly, the intercept $\beta_0^{[m+1]}$ is updated by

$$\beta_0^{[m+1]} = \beta_0^{[m]} - \left(\frac{\partial^2 l_2^{[m]}}{\partial\left(\beta_0^{[m]}\right)^2}\right)^{-1} \frac{\partial l_2^{[m]}}{\partial \beta_0^{[m]}}. \tag{17}$$

For $\boldsymbol{\beta}_j^{[m+1]} \in R^{p_j}$, consider the optimization function:

$$-\left(\boldsymbol{\beta}_j^{[m+1]} - \boldsymbol{\beta}_j^{[m]}\right)^{\mathrm{T}}\left(\frac{\partial l_2^{[m]}}{\partial \boldsymbol{\beta}_j^{[m]}} + \mu_2 V_{2j}^{[m]}\right) + \frac{M_{2j}^{[m]}}{2}\left(\boldsymbol{\beta}_j^{[m+1]} - \boldsymbol{\beta}_j^{[m]}\right)^{\mathrm{T}}\left(\boldsymbol{\beta}_j^{[m+1]} - \boldsymbol{\beta}_j^{[m]}\right) + \mu_1\sqrt{p_j}\left\|\boldsymbol{\beta}_j^{[m+1]}\right\|. \tag{18}$$

Here, $V_{2j}^{[m]}$ is p_j-length vector, and $M_{2j}^{[m]}$ is a constant as follows:

$$V_{2j}^{[m]} = \left(\frac{1}{\left|\beta_{jk}^{[m]}\right| + \tau} \left(\frac{\alpha_{jk}^{[m]}}{\left|\alpha_{jk}^{[m]}\right| + \tau} - \frac{\beta_{jk}^{[m]}}{\left|\beta_{jk}^{[m]}\right| + \tau} \right) \right)_{1 \le k \le p_j},$$
$$M_{2j}^{[m]} = \psi\left(-\left(\frac{\partial^2 l_2^{[m]}}{\partial \beta_{jk_1}^{[m]} \partial \beta_{jk_2}^{[m]}} \right)_{1 \le k_1, k_2 \le p_j} \right) + \max_k \left(\mu_2 \left(\frac{1}{\left|\beta_{jk}^{[m]}\right| + \tau} \right)^2 \right). \tag{19}$$

The tuning parameters, μ_1 and μ_2, are selected by 5-fold cross-validation. The parameter τ in the approximation of the sign function controls the degree of approximation [24]. A smaller τ leads to a better approximation but less stable estimation. The proposed method is valid as long as τ is not too large, and the parameters with different signs can be distinguished [34]. Therefore, as suggested in [31], we set $\tau = 0.1$, which leads to satisfactory results.

The ECD algorithm is summarized in Table 1.

4. Simulations

In this section, some experiment examples are given to illustrate the performance of the proposed method compared to alternative methods. The proposed method is a mixture cure model with group lasso and sign consistency (MCGS). Two alternative methods are the standard mixture cure model without variable selection and sign consistency penalty (Full) and the mixture cure model with group lasso penalty (MCG), respectively. For comparison, alternative methods both adopt the logistic regression in the incident part and the exponential model in the latency part.

Here, we set sample size $n = 1000$ and consider low-dimensional data with $p = 60$ and high-dimensional data with $p = 200$. The censoring time is generated from an exponential distribution with censoring rates $\eta = \{0.25, 5\}$. We consider three examples regarding different grouping structures of coefficients and different types of variables. The true values of coefficients are generated according to the following settings in three examples:

Example 1: for each subgroup, we set $p_j = 10$. Intragroup variables x_{jk_m} and x_{jk_n} are generated from a multivariate normal distribution with the correlation coefficient $\rho = 0.3^{|k_m - k_n|}$, whereas intergroup variables are independent. Denote the true coefficients as $\boldsymbol{\alpha}^{\text{true}}$ and $\boldsymbol{\beta}^{\text{true}}$. The coefficients of the two scenarios are shown as follows:

Scenario 1:

$$\alpha^{\text{true}} = (\underbrace{0.1, \ldots, 0.1}_{10}, \underbrace{-0.4, \ldots, -0.4}_{10}, \underbrace{0, \ldots, 0}_{p-20})^{\text{T}},$$
$$\beta^{\text{true}} = (\underbrace{0.4, \ldots, 0.4}_{10}, \underbrace{-0.3, \ldots, -0.3}_{10}, \underbrace{0, \ldots, 0}_{p-20})^{\text{T}}. \tag{20}$$

Scenario 2:

$$\alpha^{\text{true}} = (\underbrace{0.5, \ldots, 0.5}_{10}, \underbrace{0.2, \ldots, 0.2}_{8}, -0.2, -0.2, \underbrace{0, \ldots, 0}_{p-20})^{\text{T}},$$
$$\beta^{\text{true}} = (\underbrace{0.4, \ldots, 0.4}_{10}, \underbrace{-0.3, \ldots, -0.3}_{8}, \underbrace{0, \ldots, 0}_{p-20})^{\text{T}}. \tag{21}$$

Example 2: the settings are similar to Example 1 except for the subgroup settings. We set 15 variables in the first subgroup, 5 variables in the second subgroup, and 10 variables for the remaining subgroups. The coefficients are shown as follows:

$$\alpha^{\text{true}} = (\underbrace{0.5, \ldots, 0.5}_{15}, \underbrace{-0.4, \ldots, -0.4}_{5}, \underbrace{0, \ldots, 0}_{p-20})^{\text{T}},$$
$$\beta^{\text{true}} = (\underbrace{0.4, \ldots, 0.4}_{15}, \underbrace{-0.3, \ldots, -0.3}_{5}, \underbrace{0, \ldots, 0}_{p-20})^{\text{T}}. \tag{22}$$

Example 3: consider the case with some discrete variables. For each subgroup, we set $p_j = 10$. A latent variable Z_j is generated from a multivariate normal distribution with the intragroup correlation coefficient $\rho = 0.3^{|k_m - k_n|}$ and intergroup correlation independent. The coefficient setting is the same as Example 2. x_{jk} is defined as follows:

$$x_{jk} = \begin{cases} I(Z_j > 0) & j \le \dfrac{p}{2}, \\ Z_j & j > \dfrac{p}{2}. \end{cases} \tag{23}$$

The performance of each model is me asured by 5 measures. Denote $\theta \in \{\boldsymbol{\alpha}, \boldsymbol{\beta}, (\boldsymbol{\alpha}^{\text{T}}, \boldsymbol{\beta}^{\text{T}})^{\text{T}}\}$, $\widehat{\theta}$ as the estimation of θ, $\widehat{\pi}$ as the estimation of π, and $\widehat{\lambda}$ as the estimation of λ. The true positive rate (TPR), false positive rate (FPR), and mean square error (MSE) with respect to $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $(\boldsymbol{\alpha}^{\text{T}}, \boldsymbol{\beta}^{\text{T}})^{\text{T}}$ can be written as follows:

$$\text{TPR}(\theta) = \frac{\text{TP}}{\text{TP} + \text{FN}},$$
$$\text{FPR}(\theta) = \frac{\text{FP}}{\text{TN} + \text{FP}}, \tag{24}$$
$$\text{MSE}(\theta) = \frac{\sum_{j=1}^{p} \left(\theta_j^{\text{true}} - \widehat{\theta}_j\right)^2}{\sum_{j=1}^{p} \left(\theta_j^{\text{true}}\right)^2},$$

where

$$\begin{aligned} \mathrm{TP} &= \sum_{j=1}^{p} I\left(\theta_j^{\mathrm{true}} \neq 0 \cap \widehat{\theta}_j \neq 0\right), \\ \mathrm{TP} + \mathrm{FN} &= \sum_{j=1}^{p} I\left(\theta_j^{\mathrm{true}} \neq 0\right), \\ \mathrm{FP} &= \sum_{j=1}^{p} I\left(\theta_j^{\mathrm{true}} = 0 \cap \widehat{\theta}_j \neq 0\right), \\ \mathrm{TN} + \mathrm{FP} &= \sum_{j=1}^{p} I\left(\theta_j^{\mathrm{true}} = 0\right). \end{aligned} \tag{25}$$

The relative root mean square error of the cure rate estimation (RMSE_π) and the relative root mean square error of the hazard function estimation (RMSE_λ) are

$$\begin{aligned} \mathrm{RMSE}_\pi &= \frac{\sum_{i=1}^{n} \left(\widehat{\pi}_i - \pi_i\right)^2}{\sum_{i=1}^{n} \pi_i^2}, \\ \mathrm{RMSE}_\lambda &= \frac{\sum_{i=1}^{n} \left(\widehat{\lambda}_i - \lambda_i\right)^2}{\sum_{i=1}^{n} \lambda_i^2}. \end{aligned} \tag{26}$$

Tables 2–5 show the mean TPR, FPR, and MSE of the coefficients, as well as the standard deviations over the 100 replicates for each example.

As indicated in Tables 2–5, the two group selection methods (MCGS and MCG) perform significantly better than the Full method. This is expected since the group lasso can select important subgroups of variables. Comparing the two group methods, the proposed method has competitive performance compared with the MCG method. It indicates that promoting sign consistency improves the performance in terms of estimation. For instance, under Scenario 1 in Example 1 with $\eta = 5$ and $p = 200$ in Table 2, the mean MSEs of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $(\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$ for the proposed method are 0.12, 0.04, and 0.09, respectively, compared to 1.00, 0.05, and 0.63 for the MCG method and 17.3, 2.38, and 11.56 for the Full method.

Tables 6–8 show the mean RMSE of π and λ, as well as the standard deviations over the 100 replicates for each example. The results illustrate the performance in terms of prediction of the probability of nondefault and survival.

As shown in Tables 6–8 the prediction performance of group selection methods is significantly better than that of the Full method, and the proposed method has competitive performance compared with the MCG method. For example, in Example 2 with $\eta = 0.25$ and $p = 200$ in Table 8, the mean RMSE_π and RMSE_λ for the proposed method are 0.01, and 0.07, respectively, compared to 0.04 and 0.12 for the MCG method and 0.27 and 10.01 for the Full method. In addition, compared with the results regarding low- and high-dimensional settings, the group selection methods have greater advantages in prediction performance when the dimensionality is higher.

Results of simulation reveal that the proposed variable effect consistency mixture cure model can improve the performance in terms of estimation and prediction compared with alternative methods.

5. Empirical Study

In this section, we applied our proposed method to real data on credit loans. The data come from the personal loan department of a city commercial bank in China, which contains 4796 personal loan samples from 2014 to 2019 after preprocessing. The data include mortgage loans and credit loans, covering consumer durables, personal housing decoration loans, and other personal consumption loans. Censoring time C_i is the interval between the loan value date and either default or the end of observation (June 1, 2019). Therefore, censoring times of customers vary from individual to individual. It has a mean of 1.93 years and a standard deviation of 0.8. Customers whose time to default Y_i is longer than the censoring time C_i are censored ($\delta_i = 0$). 47 out of 4796 customers are censored. By transforming the discrete variables into dummy variables, the data contain 27 variables. Table 9 provides a list of variables and their descriptions.

In this section, the alternative method is a mixture cure model with group lasso (MCG). Different from the simulation, the real values of parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are unknown in real data. Referring to [30, 35], we adopt the (1) log-rank statistics and (2) negative log-likelihood to evaluate the performance of the models instead of $\mathrm{TPR}(\theta)$, $\mathrm{FPR}(\theta)$, $\mathrm{ME}(\theta)$, RMSE_π, and RMSE_λ in simulation.

Log-rank statistics is a commonly used indicator in survival analysis to test the null hypothesis that there is no significant difference in survival distribution between two or more independent groups. It is calculated by cross-validation. We sequentially take 1/10 samples as the validation set and the remaining as the training set. Apply the proposed and alternative methods to obtain the estimation of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and then calculate the $\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}$ and $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$ for the validation set. Results of $\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}$ and $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$ are based on 10 replicates. Divide the calculated $\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}$ into two groups at the median and calculate the log-rank statistics. Similarly, divide the calculated $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$ into two groups at the median and calculate the log-rank statistics. The mean log-rank of the proposed and the alternative methods is 5.6 and 4.3 respectively, indicating better performance of the proposed method.

Figure 1 shows the Kaplan–Meier curves stratified by different groups. Kaplan–Meier curves are commonly used to describe the change of the survival probability overtime in survival analysis [36, 37]. The probability of being cured is negatively related to $\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}$, and the survival time is negatively related to $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$. In Figure 1, a group is denoted by "low risk" with lower $\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}$ (a) or lower $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$ (b) whereas another group is denoted by "high risk." As indicated in Figure 1, there are clearly different trends in the curves in different groups. Customers with lower $\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}$ and $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$ have lower risk and are less likely to default.

To assess the performance of the model, the data are randomly divided into training set and test set by 2 : 1. The training set is used for fitting the model and the test set is used for evaluating the prediction performance of the fitted model. The tuning parameters are selected by 5-fold cross-validation. The mean (standard deviation) negative log-likelihood of the proposed method (MCGS) and the alternative method (MCG) is 106.04 (16.04) and 118.60

Table 1: ECD algorithm.

Expectation Coordinate Descent algorithm
Initialize $m = 0$, $\alpha_0^{[m]} = \beta_0^{[m]} = 0$, and $\alpha^{[m]} = \beta^{[m]} = 0_{J\times p}$. Compute $\pi_i^{[m]}$ and $\lambda_i^{[m]}$ from (9) and (11)
Repeat
E-step
Update $E(y_i)$ from (10)
CD-step
Update $\boldsymbol{\alpha}_0^{[m+1]}$ from (14)
For $j = 1, \ldots, p$, update $\boldsymbol{\alpha}_j^{[m+1]}$ from (15); update $\pi_i^{[m+1]}$ from (9)
Update $\boldsymbol{\beta}_0^{[m+1]}$ from (17);
For $j = 1, \ldots, p$, update $\boldsymbol{\beta}_j^{[m+1]}$ from (18); update $\lambda_i^{[m+1]}$ from (11)
$m = m + 1$
Until $\max\left\{\left\|\boldsymbol{\alpha}_j^{[m+1]} - \boldsymbol{\alpha}_j^{[m]}\right\|, \left\|\boldsymbol{\beta}_j^{[m+1]} - \boldsymbol{\beta}_j^{[m]}\right\|\right\} \le 10^{-3}$

Table 2: Results of the estimation performance in scenario 1, Example 1.

η	p		$\boldsymbol{\alpha}$			$\boldsymbol{\beta}$			$(\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$		
			Full	MCG	MCGS	Full	MCG	MCGS	Full	MCG	MCGS
0.25	60	MSE	0.65 (0.23)	0.13 (0.04)	0.07 (0.02)	0.13 (0.03)	0.04 (0.01)	0.03 (0.01)	0.34 (0.10)	0.07 (0.02)	0.05 (0.01)
		TPR	—	0.92 (0.14)	1.00 (0.01)	—	1.00 (0.00)	1.00 (0.00)	—	0.96 (0.07)	1.00 (0.00)
		FPR	—	0.18 (0.18)	0.28 (0.19)	—	0.59 (0.16)	0.48 (0.18)	—	0.38 (0.13)	0.38 (0.17)
	200	MSE	29.04 (9.28)	0.17 (0.04)	0.08 (0.02)	1.05 (0.18)	0.04 (0.01)	0.04 (0.01)	12.38 (3.76)	0.09 (0.02)	0.06 (0.01)
		TPR	—	0.83 (0.20)	1.00 (0.00)	—	1.00 (0.00)	1.00 (0.00)	—	0.91 (0.10)	1.00 (0.00)
		FPR	—	0.02 (0.03)	0.07 (0.05)	—	0.28 (0.09)	0.20 (0.09)	—	0.15 (0.05)	0.14 (0.07)
5	60	MSE	9.17 (3.02)	0.42 (0.21)	0.23 (0.04)	0.51 (0.14)	0.07 (0.02)	0.07 (0.02)	4.01 (1.23)	0.21 (0.09)	0.13 (0.02)
		TPR	—	0.55 (0.24)	1.00 (0.00)	—	1.00 (0.00)	1.00 (0.00)	—	0.77 (0.12)	1.00 (0.00)
		FPR	—	0.03 (0.08)	0.04 (0.07)	—	0.67 (0.15)	0.06 (0.11)	—	0.35 (0.09)	0.05 (0.09)
	200	MSE	17.30 (2.12)	1.00 (0.00)	0.12 (0.02)	2.38 (0.48)	0.05 (0.01)	0.04 (0.01)	11.56 (1.26)	0.63 (0.00)	0.09 (0.02)
		TPR	—	0.00 (0.00)	1.00 (0.00)	—	1.00 (0.00)	1.00 (0.00)	—	0.50 (0.00)	1.00 (0.00)
		FPR	—	0.00 (0.00)	0.02 (0.01)	—	0.35 (0.08)	0.02 (0.01)	—	0.17 (0.04)	0.02 (0.01)

Note. In each cell, there is mean (standard deviation).

(19.09), respectively. The result is based on 100 duplicates. It indicates that the proposed method performs better than the alternative in terms of model fit and prediction.

Table 10 reports the estimations of the MCGS method and the MCG method. A positive coefficient $\boldsymbol{\alpha}$ indicates that the variable is positively related to the probability of default, and a positive coefficient $\boldsymbol{\beta}$ indicates that the variable is negatively related to default time. Both probabilities of default and default time are two quite relevant credit aspects. Compared with the alternative method, the signs of the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ of the proposed method are promoted to be more consistent, whereas the business type in the MCG model has an opposite effect on the probability and time to default. The results show that promoting variable effect consistency can improve prediction performance as well as interpretability.

The coefficient results of the proposed method reveal that interest rate, loan line, business type, gender, education, and employment status are important variables that affect the probability of default and the time to default. Loan term, age, medical insurance, entrusted payment, early repayment, annual household income, type of workplace, and housing status have no significant impact on credit. The impact of occupation and professional title on credit is not clear.

From the perspective of loan products, we find that interest rate has a negative impact. This is not surprising, as higher interest rates lead to higher costs, and the customers are more likely to default. The loan line has a positive effect. One possible explanation is that low-risk customers are more likely to obtain a higher loan line. Loan term, entrusted payment, and early repayment have no effect on the credit.

Table 3: Results of the estimation performance in scenario 2, Example 1.

η	p		$\boldsymbol{\alpha}$			$\boldsymbol{\beta}$			$(\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$		
			Full	MCG	MCGS	Full	MCG	MCGS	Full	MCG	MCGS
0.25	60	MSE	0.51 (0.19)	0.11 (0.03)	0.11 (0.03)	0.13 (0.03)	0.04 (0.01)	0.04 (0.01)	0.33 (0.11)	0.08 (0.02)	0.08 (0.02)
		TPR	—	0.99 (0.03)	0.99 (0.03)	—	1.00 (0.00)	1.00 (0.00)	—	0.99 (0.01)	0.99 (0.01)
		FPR	—	0.33 (0.23)	0.33 (0.23)	—	0.71 (0.13)	0.71 (0.13)	—	0.52 (0.14)	0.52 (0.14)
	200	MSE	24.30 (6.60)	0.16 (0.03)	0.16 (0.03)	1.01 (0.19)	0.05 (0.01)	0.05 (0.01)	13.40 (3.50)	0.11 (0.02)	0.11 (0.02)
		TPR	—	0.98 (0.03)	0.98 (0.03)	—	1.00 (0.00)	1.00 (0.00)	—	0.99 (0.02)	0.99 (0.02)
		FPR	—	0.04 (0.05)	0.04 (0.05)	—	0.35 (0.08)	0.35 (0.08)	—	0.19 (0.06)	0.19 (0.06)
5	60	MSE	5.56 (1.83)	0.32 (0.10)	0.33 (0.07)	0.53 (0.15)	0.10 (0.03)	0.11 (0.03)	3.21 (0.97)	0.22 (0.06)	0.23 (0.04)
		TPR	—	0.63 (0.19)	0.87 (0.20)	—	1.00 (0.01)	1.00 (0.01)	—	0.81 (0.10)	0.93 (0.10)
		FPR	—	0.07 (0.13)	0.08 (0.11)	—	0.70 (0.15)	0.36 (0.33)	—	0.38 (0.11)	0.22 (0.19)
	200	MSE	24.46 (3.45)	0.60 (0.25)	0.39 (0.04)	4.52 (0.85)	0.11 (0.03)	0.13 (0.03)	15.13 (1.71)	0.37 (0.14)	0.27 (0.02)
		TPR	—	0.38 (0.21)	0.93 (0.17)	—	1.00 (0.00)	1.00 (0.02)	—	0.69 (0.11)	0.96 (0.08)
		FPR	—	0.00 (0.00)	0.02 (0.04)	—	0.32 (0.12)	0.08 (0.13)	—	0.16 (0.06)	0.05 (0.08)

Note. In each cell, there is mean (standard deviation).

Table 4: Results of the estimation performance in Example 2.

η	p		$\boldsymbol{\alpha}$			$\boldsymbol{\beta}$			$(\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$		
			Full	MCG	MCGS	Full	MCG	MCGS	Full	MCG	MCGS
0.25	60	MSE	0.35 (0.13)	0.13 (0.03)	0.07 (0.02)	0.09 (0.02)	0.04 (0.01)	0.02 (0.01)	0.25 (0.08)	0.09 (0.02)	0.05 (0.01)
		TPR	—	1.00 (0.00)	1.00 (0.00)	—	1.00 (0.00)	1.00 (0.00)	—	1.00 (0.00)	1.00 (0.00)
		FPR	—	0.37 (0.24)	0.08 (0.06)	—	0.76 (0.09)	0.12 (0.07)	—	0.57 (0.13)	0.10 (0.06)
	200	MSE	20.24 (4.57)	0.23 (0.05)	0.09 (0.02)	0.68 (0.12)	0.04 (0.01)	0.02 (0.01)	12.70 (2.81)	0.16 (0.03)	0.06 (0.01)
		TPR	—	1.00 (0.00)	1.00 (0.00)	—	1.00 (0.00)	1.00 (0.00)	—	1.00 (0.00)	1.00 (0.00)
		FPR	—	0.02 (0.04)	0.04 (0.02)	—	0.37 (0.09)	0.06 (0.03)	—	0.20 (0.05)	0.05 (0.02)
5	60	MSE	4.72 (1.08)	0.98 (0.10)	0.10 (0.02)	0.30 (0.07)	0.04 (0.01)	0.03 (0.01)	3.02 (0.66)	0.62 (0.06)	0.07 (0.02)
		TPR	—	0.05 (0.22)	1.00 (0.00)	—	1.00 (0.00)	1.00 (0.00)	—	0.53 (0.11)	1.00 (0.00)
		FPR	—	0.00 (0.00)	0.03 (0.03)	—	0.64 (0.17)	0.04 (0.03)	—	0.32 (0.09)	0.03 (0.03)
	200	MSE	17.30 (2.12)	1.00 (0.00)	0.12 (0.02)	2.38 (0.48)	0.05 (0.01)	0.04 (0.01)	11.56 (1.26)	0.63 (0.00)	0.09 (0.02)
		TPR	—	0.00 (0.00)	1.00 (0.00)	—	1.00 (0.00)	1.00 (0.00)	—	0.50 (0.00)	1.00 (0.00)
		FPR	—	0.00 (0.00)	0.02 (0.01)	—	0.35 (0.08)	0.02 (0.01)	—	0.17 (0.04)	0.02 (0.01)

Note. In each cell, there is mean (standard deviation).

Table 5: Results of the estimation performance in Example 3.

η	p		$\boldsymbol{\alpha}$			$\boldsymbol{\beta}$			$(\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$		
			Full	MCG	MCGS	Full	MCG	MCGS	Full	MCG	MCGS
0.25	60	MSE	0.33 (0.08)	0.19 (0.05)	0.10 (0.02)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.22 (0.05)	0.13 (0.03)	0.07 (0.01)
		TPR	—	1.00 (0.01)	1.00 (0.00)	—	1.00 (0.00)	1.00 (0.00)	—	1.00 (0.00)	1.00 (0.00)
		FPR	—	0.53 (0.15)	0.29 (0.12)	—	0.66 (0.08)	0.29 (0.12)	—	0.60 (0.09)	0.29 (0.12)
	200	MSE	2.42 (0.57)	0.48 (0.14)	0.17 (0.02)	0.20 (0.03)	0.05 (0.01)	0.03 (0.01)	1.56 (0.35)	0.31 (0.08)	0.11 (0.01)
		TPR	—	0.95 (0.18)	1.00 (0.00)	—	1.00 (0.00)	1.00 (0.00)	—	0.98 (0.09)	1.00 (0.00)
		FPR	—	0.05 (0.07)	0.14 (0.06)	—	0.57 (0.10)	0.16 (0.06)	—	0.31 (0.07)	0.15 (0.06)
5	60	MSE	1.31 (0.33)	0.99 (0.08)	0.43 (0.29)	0.08 (0.02)	0.03 (0.01)	0.03 (0.01)	0.83 (0.20)	0.62 (0.05)	0.27 (0.18)
		TPR	—	0.01 (0.10)	0.83 (0.38)	—	1.00 (0.00)	1.00 (0.00)	—	0.51 (0.05)	0.92 (0.19)
		FPR	—	0.00 (0.02)	0.14 (0.14)	—	0.51 (0.13)	0.36 (0.17)	—	0.25 (0.07)	0.25 (0.12)
	200	MSE	6.94 (1.12)	1.00 (0.00)	0.28 (0.06)	0.51 (0.07)	0.04 (0.01)	0.05 (0.01)	4.47 (0.68)	0.63 (0.00)	0.19 (0.04)
		TPR	—	0.00 (0.00)	1.00 (0.00)	—	1.00 (0.00)	1.00 (0.00)	—	0.50 (0.00)	1.00 (0.00)
		FPR	—	0.00 (0.00)	0.17 (0.06)	—	0.34 (0.08)	0.22 (0.06)	—	0.17 (0.04)	0.19 (0.06)

Note. In each cell, there is mean (standard deviation).

Table 6: Results of the prediction performance in Example 1.

η	p	RMSE_{π}			RMSE_{λ}		
		Full	MCG	MCGS	Full	MCG	MCGS
				Scenario 1			
0.25	60	0.06 (0.02)	0.02 (0.01)	0.01 (0.00)	0.57 (0.80)	0.10 (0.08)	0.12 (0.08)
	200	0.37 (0.04)	0.03 (0.01)	0.02 (0.01)	24.10 (57.28)	0.12 (0.08)	0.16 (0.10)
5	60	0.33 (0.05)	0.19 (0.18)	0.10 (0.03)	6.32 (20.47)	0.23 (0.28)	0.18 (0.09)
	20	0.41 (0.03)	0.96 (0.04)	0.02 (0.01)	>10^3 (>10^3)	0.15 (0.15)	0.10 (0.08)
				Scenario 2			
0.25	60	0.06 (0.01)	0.02 (0.01)	0.02 (0.01)	0.52 (0.71)	0.12 (0.07)	0.12 (0.07)
	200	0.31 (0.03)	0.03 (0.01)	0.03 (0.01)	72.58 (387.68)	0.15 (0.10)	0.15 (0.10)
5	60	0.30 (0.06)	0.08 (0.04)	0.14 (0.06)	5.72 (10.42)	0.21 (0.18)	0.27 (0.15)
	200	0.53 (0.04)	0.26 (0.24)	0.17 (0.04)	>10^3 (>10^3)	0.26 (0.17)	0.35 (0.17)

Note. In each cell, there is mean (standard deviation).

Different coefficient of business type reveals that, compared with other personal loans, consumer durables are more likely to default.

From the perspective of the influence of the variables of the customers, employed customers have a positive impact. Age, annual household income, housing status, and type of workplace have no significant effect on the credit. Compared with women, men are more likely to default. This is consistent with the results of [38] and the personality characteristics of men's risk preference [39]. Customers with higher education are less likely to default. Bachelor degree or above has a positive effect on credit. Generally, customers

Table 7: Results of the prediction performance in Example 2.

η	p	$RMSE_\pi$			$RMSE_\lambda$		
		Full	MCG	MCGS	Full	MCG	MCGS
0.25	60	0.05 (0.01)	0.02 (0.01)	0.01 (0.00)	0.36 (0.35)	0.10 (0.07)	0.06 (0.05)
	200	0.27 (0.03)	0.04 (0.01)	0.01 (0.00)	10.01 (33.80)	0.12 (0.17)	0.07 (0.06)
5	60	0.26 (0.03)	0.88 (0.16)	0.02 (0.01)	1.98 (4.56)	0.21 (0.30)	0.09 (0.08)
	200	0.41 (0.03)	0.96 (0.04)	0.02 (0.01)	>10^3 (>10^3)	0.15 (0.15)	0.10 (0.08)

Note. In each cell, there is mean (standard deviation).

Table 8: Results of the prediction performance in Example 3.

η	p	$RMSE_\pi$			$RMSE_\lambda$		
		Full	MCG	MCGS	Full	MCG	MCGS
0.25	60	0.21 (0.04)	0.11 (0.03)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)
	200	0.83 (0.10)	0.28 (0.08)	0.10 (0.02)	0.30 (0.07)	0.05 (0.01)	0.03 (0.01)
5	60	1.00 (0.19)	0.77 (0.09)	0.34 (0.21)	0.09 (0.03)	0.03 (0.01)	0.03 (0.01)
	200	2.62 (0.21)	0.72 (0.06)	0.21 (0.10)	0.99 (0.33)	0.04 (0.01)	0.04 (0.01)

Note. In each cell, there is mean (standard deviation).

Table 9: Description of variables.

Variables	Description
Business type	Consumer durables, personal housing decoration loans, and other personal consumption loans
Interest rate	$[0.029, 0.095]$
Loan line	$(0, +\infty)$
Loan term	$(0, +\infty)$
Early repayment	Yes, no
Entrusted payment	Yes, no
Age	$[18, 70]$
Gender	Male, female
Education	Master/doctor, bachelor, vocational education, high school, and below
Medical insurance	Yes, no
Housing status	Self-purchasing (with mortgage), self-purchasing (without mortgage), and others
Employment	Employed and others
Type of workplace	Government organization/institution, firm, and others
Occupation	Managers, commercial and service workers, and others
Professional title	Advanced, intermediate, primary, and no professional title
Annual household income (RMB)	$\leq 200,000$, $200,000 - 400,000$, $400,000 - 600,000$, $\geq 600,000$, and unknown

with higher education have a higher chance of getting decent jobs and income. They tend to maintain good credit records and are less likely to default. Compared with other employment groups such as self-employed, freelance, and unemployed ones, the employed group has a more stable income and is less likely to default.

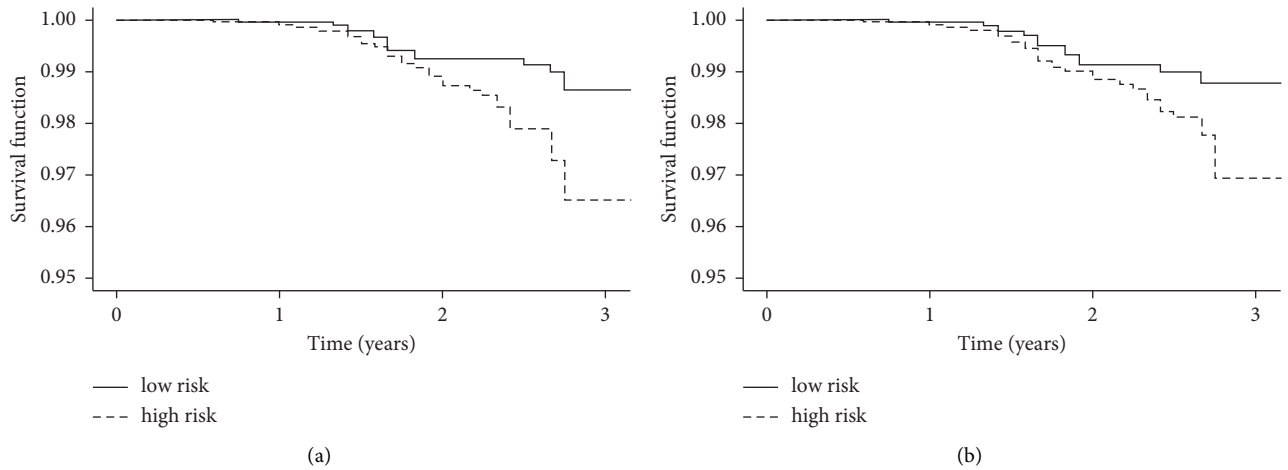

Figure 1: Kaplan–Meier curves stratified by different groups: (a) divided by the median of $\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}$ and (b) divided by the median of $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$.

Table 10: Estimation of coefficients of the methods.

	MCG		MCGS	
	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$
$\boldsymbol{\alpha}_0/\boldsymbol{\beta}_0$	−1.17	−4.07	−1.43	−3.67
Business type				
Consumer durables	0.15	0.70	0.11	0.12
Personal housing decoration loans	−0.02	0.21	0	0
Interest rate	0.54	0.17	0.34	0.28
Loan line	−0.06	0	−0.03	−0.03
Loan term	0	0	0	0
Early repayment (yes)	0	−0.24	0	0
Entrusted payment (yes)	−0.13	0	0	0
Age	0	0	0	0
Gender (male)	0	0.82	0.25	0.30
Education				
Master/doctor	−0.03	−0.32	−0.02	−0.02
Bachelor	−0.04	−0.22	−0.06	−0.06
Vocational education	0.02	0.32	0.05	0.05
Medical insurance (yes)	0	0.34	0	0
Housing status				
Self-purchasing (with mortgage)	0	−0.01	0	0
Self-purchasing (without mortgage)	0	0.04	0	0
Employment (employed)	−0.08	−0.31	−0.24	−0.21
Type of workplace				
Government organization and institution	−0.14	−1.53	0	0
Firm	−0.14	−0.40	0	0
Occupation				
Managers	0.09	0.09	0.04	0.04
Commercial and service workers	−0.25	−0.29	−0.11	−0.11
Professional title				
Advanced	0.04	0.41	0.04	0.04
Intermediate	−0.02	−0.25	−0.04	−0.04
Primary	0	−0.03	0	0
Annual household income (RMB)				
≤200,000	0	0.05	0	0
200,000 – 400,000	0	0.38	0	0
400,000 – 600,000	0	−0.29	0	0
≥600,000	0	−0.19	0	0

6. Conclusions

The mixture cure model is widely adopted in credit scoring for its ability of predicting whether customers will default and when they are likely to default. However, most of the existing studies ignore the relations between the two sets of variable effects in the two model parts which may get conflicting results of variable effects. It can be difficult to interpret the results and apply them in practice.

In this paper, we propose a variable effect consistency mixture cure model, to promote the similarity of the sign of variables in the two model parts by imposing a sign consistency penalty. Meanwhile, to accommodate the high-dimensional credit data, we also impose a group lasso penalty to conduct variable selection and parameter estimation. Simulation shows that the proposed method has competitive performance compared with the MCG method and significantly outperforms the Full method in terms of estimation and prediction. Furthermore, the empirical study illustrates that the proposed method can improve prediction performance as well as interpretability. The results of the variable effect consistency mixture cure model also offer additional insights into the relationship between the variable effect before and after loan.

Data Availability

The raw/processed data used in the empirical study cannot be shared at this time as the data also form part of an ongoing study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the National Office for Philosophy and Social Sciences of China under Grant no. 20&ZD137 and the National Bureau of Statistics of China under Grant no. 2020ZX20.

References

[1] D. M. B. Silva, G. H. A. Pereira, and T. M. Magalhães, "A class of categorization methods for credit scoring models," *European Journal of Operational Research*, vol. 296, no. 1, pp. 323–331, 2022.

[2] P. Pławiak, M. Abdar, J. Pławiak, V. Makarenkov, and U. R. Acharya, "A new deep genetic hierarchical network of learners for prediction of credit scoring," *Information Sciences*, vol. 516, pp. 401–418, 2020.

[3] R. Y. Goh, L. S. Lee, H.-V. Seow, and K. Gopal, "Hybrid harmony search–artificial intelligence models in credit scoring," *Entropy*, vol. 22, no. 9, https://doi.org/10.3390/e22090989, Article ID 989, 2020.

[4] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen, "Benchmarking state-of-the-art classification algorithms for credit scoring," *Journal of the Operational Research Society*, vol. 54, no. 6, pp. 627–635, 2003.

[5] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research," *European Journal of Operational Research*, vol. 247, no. 1, pp. 124–136, 2015.

[6] Q. Zhou, L. Wang, L. Juan, S. Zhou, and L. Li, "The study on credit risk warning of regional listed companies in China based on logistic model," *Discrete Dynamics in Nature and Society*, vol. 2021, Article ID 6672146, 8 pages, 2021.

[7] L. Zhang, H. Ray, J. Priestley, and S. Tan, "A descriptive study of variable discretization and cost-sensitive logistic regression on imbalanced credit data," *Journal of Applied Statistics*, vol. 47, no. 3, pp. 568–581, 2020.

[8] S. Maldonado, G. Peters, and R. Weber, "Credit scoring using three-way decisions with probabilistic rough sets," *Information Sciences*, vol. 507, pp. 700–714, 2020.

[9] C. Lohmann and T. Ohliger, "Using accounting-based and loan-related information to estimate the cure probability of a defaulted company," *European Financial Management*, vol. 27, pp. 620–640, 2021, https://doi.org/10.1111/eufm.12279.

[10] B. Narain, "Survival Analysis and the Credit Granting Decision," in *Credit Scoring And Credit Control*, pp. 109–121, Oxford University Press, New York, NY, USA, 1992.

[11] J. Banasik, J. N. Crook, and L. C. Thomas, "Not if but when will borrowers default," *Journal of the Operational Research Society*, vol. 50, no. 12, pp. 1185–1190, 1999.

[12] M. Stepanova and L. Thomas, "Survival analysis methods for personal loan data," *Operations Research*, vol. 50, no. 2, pp. 277–289, 2002.

[13] T. Bellotti and J. Crook, "Credit scoring with macroeconomic variables using survival analysis," *Journal of the Operational Research Society*, vol. 60, no. 12, pp. 1699–1707, 2009.

[14] V. B. Djeundje and J. Crook, "Dynamic survival models with varying coefficients for credit risks," *European Journal of Operational Research*, vol. 275, no. 1, pp. 319–333, 2019.

[15] Y. Xia, L. He, Y. Li, Y. Fu, and Y. Xu, "A dynamic credit scoring model based on survival gradient boosting decision tree approach," *Technological and Economic Development of Economy*, vol. 27, no. 1, pp. 96–119, 2021, https://doi.org/10.3846/tede.2020.13997.

[16] E. N. C. Tong, C. Mues, and L. C. Thomas, "Mixture cure models in credit scoring: if and when borrowers default," *European Journal of Operational Research*, vol. 218, no. 1, pp. 132–139, 2012.

[17] L. Dirick, G. Claeskens, and B. Baesens, "Time to default in credit scoring using survival analysis: a benchmark study," *Journal of the Operational Research Society*, vol. 68, no. 6, pp. 652–665, 2017.

[18] B. C. Alves and J. G. Dias, "Survival mixture models in behavioral scoring," *Expert Systems with Applications*, vol. 42, no. 8, pp. 3902–3910, 2015.

[19] N. Zhang, Q. Yang, A. Kelleher, and W. Si, "A new mixture cure model under competing risks to score online consumer loans," *Quantitative Finance*, vol. 19, no. 7, pp. 1243–1253, 2019.

[20] C. Jiang, Z. Wang, and H. Zhao, "A prediction-driven mixture cure model and its application in credit scoring," *European Journal of Operational Research*, vol. 277, no. 1, pp. 20–31, 2019.

[21] L. Dirick, T. Bellotti, G. Claeskens, and B. Baesens, "Macroeconomic factors in credit risk calculations: including time-varying covariates in mixture cure models," *Journal of Business & Economic Statistics*, vol. 37, no. 1, pp. 40–53, 2019.

[22] C. Han and R. Kronmal, "Two-part models for analysis of Agatston scores with possible proportionality constraints," *Communications in Statistics-Theory and Methods*, vol. 35, no. 1, pp. 99–111, 2006.
[23] F. Liu, Z. Hua, and A. Lim, "Identifying future defaulters: a hierarchical Bayesian method," *European Journal of Operational Research*, vol. 241, no. 1, pp. 202–211, 2015.
[24] Q. Zhang, S. Ma, and Y. Huang, "Promote sign consistency in the joint estimation of precision matrices," *Computational Statistics & Data Analysis*, vol. 159, 2021 https://doi.org/10.1016/j.csda.2021.107210, Article ID 107210.
[25] Q. Zhang, S. Zhang, J. Liu, J. Huang, and S. Ma, "Penalized integrative analysis under the accelerated failure time model," *Statistica Sinica*, vol. 26, no. 2, pp. 492–508, 2016.
[26] J. Huang, P. Breheny, and S. Ma, "A selective review of group selection in high-dimensional models," *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, vol. 27, pp. 481–499, 2012.
[27] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B*, vol. 68, no. 1, pp. 49–67, 2006.
[28] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York, NY, USA, 2009.
[29] M. E. Ghitany, R. A. Maller, and S. Zhou, "Exponential mixture models with long-term survivors and covariates," *Journal of Multivariate Analysis*, vol. 49, no. 2, pp. 218–241, 1994.
[30] X. Fan, M. Liu, K. Fang, Y. Huang, and S. Ma, "Promoting structural effects of covariates in the cure rate model with penalization," *Statistical Methods in Medical Research*, vol. 26, no. 5, pp. 2078–2092, 2017.
[31] C. Zheng, J. Zhu, and J. Zhu, "Promote sign consistency in cure rate model with Weibull lifetime," *AIMS Mathematics*, vol. 7, no. 2, pp. 3186–3202, 2022.
[32] X. Shi, S. Ma, and Y. Huang, "Promoting sign consistency in the cure model estimation and selection," *Statistical Methods In Medical Research*, vol. 29, no. 1, pp. 15–28, 2020.
[33] Y. Yang and H. Zou, "A fast unified algorithm for solving group-lasso penalize learning problems," *Statistics and Computing*, vol. 25, no. 6, pp. 1129–1141, 2015.
[34] K. Fang, X. Fan, Q. Zhang, and S. Ma, "Integrative sparse principal component analysis," *Journal of Multivariate Analysis*, vol. 166, pp. 1–16, 2018.
[35] J. Rodrigues, M. de Castro, V. G. Cancho, and N. Balakrishnan, "COM-Poisson cure rate survival models and an application to a cutaneous melanoma data," *Journal of Statistical Planning and Inference*, vol. 139, no. 10, pp. 3605–3611, 2009.
[36] T. Chen and P. Du, "Promotion time cure rate model with nonparametric form of covariate effects," *Statistics in Medicine*, vol. 37, no. 10, pp. 1625–1635, 2018.
[37] S. Pal and N. Balakrishnan, "Likelihood inference based on EM algorithm for the destructive length-biased Poisson cure rate model with Weibull lifetime," *Communications in Statistics - Simulation and Computation*, vol. 47, no. 3, pp. 644–660, 2018.
[38] Y. Li, Y. Li, and Y. Li, "What factors are influencing credit card customer's default behavior in China? A study based on survival analysis," *Physica A: Statistical Mechanics and Its Applications*, vol. 526, Article ID 120861, 2019.
[39] Y. Shu and Q. Y. Yang, "Research on auto loan default prediction based on large sample data model," *Management Review*, vol. 29, no. 9, pp. 59–71, 2017.