Statistical Modelling 2024; **24(2)**: 115–138

On Lasso and adaptive Lasso for non-random sample in credit scoring

Emmanuel O. Ogundimu[1]
[1]Department of Mathematical Sciences, Durham University, Durham, UK

Abstract: Prediction models in credit scoring are often formulated using available data on accepted applicants at the loan application stage. The use of this data to estimate probability of default (PD) may lead to bias due to non-random selection from the population of applicants. That is, the PD in the general population of applicants may not be the same with the PD in the subpopulation of the accepted applicants. A prominent model for the reduction of bias in this framework is the sample selection model, but there is no consensus on its utility yet. It is unclear if the bias-variance trade-off of regularization techniques can improve the predictions of PD in non-random sample selection setting. To address this, we propose the use of Lasso and adaptive Lasso for variable selection and optimal predictive accuracy. By appealing to the least square approximation of the likelihood function of sample selection model, we optimize the resulting function subject to L_1 and adaptively weighted L_1 penalties using an efficient algorithm. We evaluate the performance of the proposed approach and competing alternatives in a simulation study and applied it to the well-known American Express credit card dataset.

Key words: Adaptive Lasso, Heckman model, reject inference, non-random selection, credit risk, bivariate copula

Received June 2021; **revised** February 2022; **accepted** February 2022

1 Introduction

Credit scoring models are used to evaluate the likelihood of credit applicants defaulting in order to decide whether to grant them credit. The scoring systems are based on the past performance of consumers who are similar to those who will be assessed under the system. In other words, several loan applicant attributes are used to assign a score. These scores are used to determine credit worthiness of the applicant. In practice, the credit scores are transformed into the probability of default (PD). PD is the expected probability that a borrower will default on the debt before its maturity. A key concern in the use of these models is that they are typically designed and calibrated using data from applicants who were previously considered adequately

Address for correspondence: Emmanuel Ogundimu, Department of Mathematical Sciences, Durham University, Durham DH1 3LE, UK
E-mail: emmanuel.ogundimu@durham.ac.uk

10.1177/1471082X221092181

creditworthy to have been granted credit (Banasik et al., 2003). Consider, as an example, where a loan application is made to a bank. The bank uses the loan applicant attributes to grant or reject the loan request. If the request is accepted, then the bank will observe the loan performance over time. Marshall et al. (2010) classified these procedures into two: the credit granting process (accept or reject) and loan performance process (default or non-default). A model developed using the accept-only applicants from the credit granting process may be a non-random sample from the target population and can lead to selection bias.

A strategy for addressing the problem of sample selection bias in credit scoring is the reject inference techniques (Hand & Henley, 1993; Crook & Banasik, 2004). Reject inference is the process of inferring how rejected loan applicants would have behaved had they been granted loan. The techniques for reject inference can be classified under two different assumptions (Kim & Sohn, 2007). The first assumption is that the distribution pattern of accepted applicants can be extended to that of rejected ones. That is, $\mathrm{P}(\text{default}|X, \text{rejected}) = \mathrm{P}(\text{default}|X, \text{accepted})$, where X is the vector of applicants' attributes. This implies that PD in the population of accepted applicants can be applied to the rejected ones. Examples of statistical methods in this category include re-weighting and extrapolation methods. The second assumption implies that $\mathrm{P}(\text{default}|X, \text{rejected}) \neq \mathrm{P}(\text{default}|X, \text{accepted})$. In this case, the PD in the population at large, $\mathrm{P}(\text{default}|X)$ cannot be approximated by the conditional model based on $\mathrm{P}(\text{default}|X, \text{accepted})$ for an applicant selected at random from the full population. A widely used method under this assumption is the bivariate probit model with sample selection (Dubin & Rivers, 1989).

There are two discordant viewpoints on the utility of sample selection models for reject inference in the literature. Greene (1998) and Greene (2008), for example, analysed the risk of a loan default for credit cardholders using sample selection model, and concluded that the model with adjustment for sample selection bias exhibits better discrimination than the model based on accept-only data. By taking variable selection into account, Marshall et al. (2010) showed that the model without considering sample selection bias can underestimate PD. Other studies that reported higher model performance can be found in Banasik et al. (2003), Banasik and Crook (2007) and Kim & Sohn (2007). On the other hand, Little (1985) and Crook & Banasik (2004) showed that adjusting for selectivity bias may not yield improved predictions when the proportion of rejected applicants is low. In a simulation study, Wu and Hand (2007) also reported the importance of the proportion of accepted or rejected applicants in reject inference. It was shown that even with the normality assumption in place for sample selection model, correction for selection bias may not improve predictions when the proportion of accepted applicants is large. Further examples can be found in Puhani (2000) and Chen and Åstebro (2012).

There are various reasons for the discordant viewpoints mentioned above. These include the proportion of rejected applicants, the inclusion of 'noise' variables (variables that are not predictive of PD) in both the loan granting and loan performance processes, which may lead to overfitting, and the degree of correlation between the error terms in loan granting and loan performance processes. Indeed, some of these

issues have been dealt with to some extent in the literature. Marshall et al. (2010) used bootstrap variable selection to control for the effect of noise variables, but their method was not optimized for predictions like regularization methods. Data mining techniques have been used in reject inference to improve the quality of credit scorecards, but these are yet to be applied to the Heckman-type selection models (see Li et al., 2017 and references therein). The need to harmonize these issues within the Heckman-type selection models for reject inference is the motivation for this article.

The contribution of this article is therefore twofold. First, we introduce Lasso (Tibshirani, 1996) and adaptive Lasso (Zou, 2006) penalized Heckman-type bivariate probit model and assess its performance in identifying predictive features of PD in credit scoring. Since the model is made up of two components, each of which may have different variables, features selection is somewhat complex. Thus, our framework appeals to the unified treatment of L_1-constrained model selection of Wang and Leng (2007), which is based on least squares approximation (LSA) of the likelihood function. The resulting LSA is then solved subject to L_1 and adaptively weighted L_1 penalties using the coordinate descent algorithm. Unlike the bootstrap variable selection approach of Marshall et al. (2010), regularization methods have the advantage of simultaneous estimation of parameters and selection of variables. Second, since variable selection provides sparse solution for the true model with true zero coefficients, the predictive performance of the model can be enhanced. We therefore propose a bootstrap internal validation method (Harrell et al., 1996; Ogundimu, 2019) for both the regularized and unregularized sample selection models. Unlike in previous work, where model validation is done using hold-out sample, the bootstrap approach can be used to quantify the degree of optimism in the model.

The remainder of the article is organized as follows. In Section 2, we describe the dataset used in the study. The bivariate probit model with sample selection (BPSSM) and its extension using copula-based sample selection model (CBSSM) are described in Section 3. We develop Lasso and adaptive Lasso estimators for the models and provide the computational algorithm for its maximization in Section 4. In Section 5, we describe five metrics for predictive performance that are not threshold dependent and the procedure for internal validation. Simulation study and data example are presented in Section 6. Finally, in Section 7, we provide concluding remarks and further results are presented in Supplementary Materials. We also provide a package in R (*HeckmanSelect*) for the implementation of the methods.

2 Dataset

We used the American Express credit card dataset (Greene, 1998, 2008) in this study. The dataset consisted of 13,444 observations on credit card applications received in a single month in 1988. Of the full sample, 10,499 applications were approved, and the next 12 months of spending and default behaviour were observed. Important variables in the data include demographic and socioeconomic factors of the applicants (e.g., Age, Income, whether the applicant owns his or her home, whether the applicant is self-employed or not and the number of dependents living with the applicant). An

Table 1 Distribution of the outcomes

(a) Distribution of events (all observations)

Event	sample
$D = 1, C = 1$	996/13444
$D = 0, C = 1$	9503/13444
$C = 0$	2945/13444

(b) Distribution of events (selected sample)

Event	sample
$D = 1$	996/10499
$D = 0$	9503/10499

important factor for granting the credit facility is grouped under 'Derogatories and Other Credit Data'. These influential variables are number of major and minor derogatory reports (60-day and 30-day delinquencies). Details of the variables used in this study can be found in Table A1 (Supplementary Materials).

The dataset consisted of two outcome variables–Cardholder status (C), which takes 1 if the application for a credit card was accepted and 0 if not, and Default status (D) which takes 1 if defaulted and 0 if not. Default is defined as having skipped payment for six months, and the corresponding status is observed only when C = 1, that is for 10,499 observations.

Table 1a shows the distribution of the cardholder status. Out of the 13,444 applicants, 2,945 (21.9%) are censored, 996 (7.41%) of those that are selected to receive the card defaulted and 9,503 (70.7%) applicants paid back their loans. Table 1b shows the default status distribution of the selected sample. As it is common in credit scoring, the event rate is less than 10%.

3 Sample selection model with binary outcome

The use of sample selection model in reject inference assumes that

$$\mathrm{P}(\text{default}|X, \text{rejected}) \neq \mathrm{P}(\text{default}|X, \text{accepted}).$$

This implies reject inference can be construed as a missing data problem under the assumption of missing not at random (MNAR) and Heckman selection model can be adapted for parameter estimation and inference. Henceforth, we treat the loan granting process (accept/reject) as the selection equation (S_i) and loan performance (default/non-default) as the outcome submodel of interest (Y_i).

Let $Y^\star$ and $S^\star$ be two latent (unobservable) variables characterizing the outcome and selection equations respectively. That is,

$$\begin{aligned} Y_i^\star &= \boldsymbol{\beta}^T \mathbf{x}_i + \varepsilon_{1i} \\ S_i^\star &= \boldsymbol{\gamma}^T \mathbf{w}_i + \varepsilon_{2i}, \quad i = 1, \ldots, n, \end{aligned} \tag{3.1}$$

where $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \ldots, \beta_p)$ and $\boldsymbol{\gamma}^T = (\gamma_0, \gamma_1, \ldots, \gamma_q)$ are unknown parameters with corresponding covariates $\mathbf{x}_i^T = (1, x_{i1}, \ldots, x_{ip})$ and $\mathbf{w}_i^T = (1, w_{i1}, \ldots, w_{iq})$; and $\varepsilon_i^T = (\varepsilon_{1i}, \varepsilon_{2i})$ are random errors with means zero, variances one and correlation ρ. Define

further Y_i and S_i as two observable versions of equation (3.1) such that

$$Y_i = \begin{cases} 1 & \text{if } Y_i^\star > 0 \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad S_i = \begin{cases} 1 & \text{if } S_i^\star > 0 \\ 0 & \text{otherwise} \end{cases}.$$

The probability mass function (PMF) of Y_i and S_i is Bernoulli, where the probability of success depends on the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ respectively.

3.1 Bivariate probit model with sample selection (BPSSM)

Suppose that the error terms in equation (3.1) follow a bivariate normal distribution

$$\begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} \sim N_2 \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right\},$$

we have the classical bivariate probit model. The selection process is such that Y_i is observed if $S_i = 1$, and Y_i is missing if $S_i = 0$. There is no selection bias when $\rho = 0$. In this case, the missing data mechanism is said to be ignorable.

Now, we have three levels of observability: $S_i = 0$ (rejected loans), $S_i = 1,\ Y_i = 0$ (accepted loans and non-default) and $S_i = 1,\ Y_i = 1$ (accepted loans and default). Thus,

$$\begin{aligned} P(S_i = 0) &= 1 - \Phi(\boldsymbol{\gamma}^T \mathbf{w}_i) = \Phi(-\boldsymbol{\gamma}^T \mathbf{w}_i) \\ P(Y_i = 0, S_i = 1) &= \Phi(\boldsymbol{\gamma}^T \mathbf{w}_i) - \Phi_2(\boldsymbol{\beta}^T \mathbf{x}_i, \boldsymbol{\gamma}^T \mathbf{w}_i; \rho) = \Phi_2(-\boldsymbol{\beta}^T \mathbf{x}_i, \boldsymbol{\gamma}^T \mathbf{w}_i; -\rho) \\ P(Y_i = 1, S_i = 1) &= \Phi_2(\boldsymbol{\beta}^T \mathbf{x}_i, \boldsymbol{\gamma}^T \mathbf{w}_i; \rho), \end{aligned} \tag{3.2}$$

where $\Phi(\cdot)$ and $\Phi_2(\cdot, \cdot; \rho)$ denote the univariate and bivariate standard normal cumulative distribution functions (CDF) respectively. The appropriate log-likelihood function is easily derived from equation (3.2) as

$$\begin{aligned} l(\boldsymbol{\theta}) = \sum_{i=1}^{n} &(1 - S_i) \ln \Phi(-\boldsymbol{\gamma}^T \mathbf{w}_i) + S_i(1 - Y_i) \ln \Phi_2(-\boldsymbol{\beta}^T \mathbf{x}_i, \boldsymbol{\gamma}^T \mathbf{w}_i; -\rho) \\ &+ S_i Y_i \ln \Phi_2(\boldsymbol{\beta}^T \mathbf{x}_i, \boldsymbol{\gamma}^T \mathbf{w}_i; \rho), \end{aligned} \tag{3.3}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \rho)$. The probability of interest is

$$P(Y_i = 1 \Big| \mathbf{x}_i, \mathbf{w}_i, S_i = 1) = \frac{\Phi_2(\boldsymbol{\beta}^T \mathbf{x}_i, \boldsymbol{\gamma}^T \mathbf{w}_i; \rho)}{\Phi(\boldsymbol{\gamma}^T \mathbf{w}_i)}. \tag{3.4}$$

It is straightforward to show that equation (3.4) reduces to $\Phi(\boldsymbol{\beta}^T\mathbf{x}_i)$ when $\rho = 0$. This is indeed the model for the accept-only applicants. That is, $\mathrm{P}(Y_i = 1|\mathbf{x}_i) = \Phi(\boldsymbol{\beta}^T\mathbf{x}_i)$, a probit regression model. The performance of the rejected loans can be imputed via

$$P(Y_i = 1\Big|\mathbf{x}_i, \mathbf{w}_i, S_i = 0) = \frac{\Phi_2(\boldsymbol{\beta}^T\mathbf{x}_i, -\boldsymbol{\gamma}^T\mathbf{w}_i; -\rho)}{\Phi(-\boldsymbol{\gamma}^T\mathbf{w}_i)}.$$

The evaluation of the performance of our method is based on equation (3.4). This is the PD given that a loan is accepted.

3.2 Copula-based sample selection model

Although there is no reason to discountenance the symmetric dependence and the underlying normal assumption used in Section 3.1 for prediction purposes, the model can be generalized using copulas. Let us define the marginal CDFs of $Y_i^\star$ and $S_i^\star$ as $F_{Y^\star}(Y_i^\star) = P(Y_i^\star \leq y_i^\star)$ and $F_{S^\star}(S_i^\star) = P(S_i^\star \leq s_i^\star)$ respectively, then their joint CDF can be written as

$$F(y_i^\star, s_i^\star|\boldsymbol{\theta}) = P(Y_i^\star \leq y_i^\star, S_i^\star \leq s_i^\star) = C\Big(F_{Y^\star}(y^\star|\boldsymbol{\beta}), F_{S^\star}(s^\star|\boldsymbol{\gamma}); \rho\Big),$$

where $C(\cdot, \cdot)$ is a two-dimensional copula function, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are as defined in equation (3.1) and ρ is an association copula parameter representing the dependence between the two marginal distributions. Note that ρ is defined as an association measure in this case and therefore can have values outside the usual correlation range of $[-1, 1]$.

Since the realized 'outcomes' Y and S are both binary, we can define the probability of event $(Y_i = 1, S_i = 1)$ as

$$p_{11i} = P(Y_i = 1, S_i = 1) = C\Big(P(Y_i = 1), P(S_i = 1); \rho\Big),$$

where

$$P(Y_i = 1) = P(Y_i^\star > 0) = 1 - F_{Y^\star}(-\boldsymbol{\beta}^T\mathbf{x}_i) \quad \text{and}$$
$$P(S_i = 1) = P(S_i^\star > 0) = 1 - F_{S^\star}(-\boldsymbol{\gamma}^T\mathbf{w}_i).$$

In addition,

$$p_{0i} = P(S_i = 0) = F_{S^\star}(-\boldsymbol{\gamma}^T\mathbf{w}_i)$$
$$p_{01i} = P(Y_i = 0, S_i = 1) = 1 - F_{S^\star}(-\boldsymbol{\gamma}^T\mathbf{w}_i) - C\Big(P(Y_i = 1), P(S_i = 1); \rho\Big).$$

Thus, the likelihood function in equation (3.3) can be generalized as

$$l(\boldsymbol{\theta}_c) = \sum_{i=1}^{n}(1 - S_i)\ln(p_{0i}) + S_i(1 - Y_i)\ln(p_{01i}) + S_i Y_i \ln(p_{11i}). \tag{3.5}$$

If we assume a Gaussian copula with normal marginals, that is $\Phi_2\Big(\Phi^{-1}\Phi(\boldsymbol{\beta}^T\mathbf{x}_i), \Phi^{-1}\Phi(\boldsymbol{\gamma}^T\mathbf{w}_i)); \rho\Big)$, then equations (3.5) and (3.3) are essentially the same. The probability of interest given in equation (3.4) is then generalized as

$$P(Y_i = 1\Big|\mathbf{x}_i, \mathbf{w}_i, S_i = 1) = \frac{P(Y_i = 1|\mathbf{x}_i, S_i = 1|\mathbf{w}_i)}{P(S_i = 1|\mathbf{w}_i)} = \frac{C\Big(P(Y_i = 1), P(S_i = 1); \rho\Big)}{1 - F_{S^\star}(-\boldsymbol{\gamma}^T\mathbf{w}_i)}.$$

Details of various copulas that can be used in non-Gaussian sample selection models, including the marginal distributions can be found in Marra et al. (2017b) and Gomes et al. (2019). We illustrate the proposed method using Ali–Mikhail–Haq (AMH) copula function with Gaussian marginal distribution for both the outcome and the selection equations. We note that AMH copula can only allow for relatively modest dependence (see Section A.3. in Supplementary Materials), and as such, we only consider comparable dependence between the outcome and the selection process of BPSSM and CBSSM in our simulation settings.

4 Lasso and adaptive Lasso

4.1 Lasso and adaptive Lasso for BPSSM and CBSSM

Ogundimu (2021) introduced a regularization method for sample selection model for continuous outcomes. We generalize the method to binary outcomes by using the unified Lasso approach of Wang and Leng (2007). Since the model is a two-component model, similar to mixture models, we adapt the method implemented in Zeng et al. (2014).

Consider the log-likelihood function $l(\boldsymbol{\theta})$ given in equation (3.3) (equivalently equation (3.5)). Suppose that the last three elements of $\boldsymbol{\theta}$ are β_0, γ_0 and ρ, and that the first $p + q$ elements of $\boldsymbol{\theta}$ are β_j, $j = 1, \ldots, p$ and γ_k, $k = 1, \ldots, q$. This construction is to ensure that β_0, γ_0 and ρ are not penalized. The Lasso estimator (Tibshirani, 1996) for the sample selection model is given by

$$\hat{\boldsymbol{\theta}}_{\text{lasso}}(\lambda) = \underset{\theta}{\text{argmin}}\left\{-l(\boldsymbol{\theta}) + \lambda\sum_{d=1}^{p+q}|\theta_d|\right\} \qquad \lambda \geq 0, \tag{4.1}$$

where the second term in the RHS of equation (4.1) is the L_1-penalty which shrinks small coefficients to zero to obtain sparse representation of the solution and λ is a tuning parameter controlling the amount of shrinkage, often chosen via cross-validation. The optimization problem reduces to the familiar maximum likelihood estimation when $\lambda = 0$.

Equation (4.1) does not have a closed form solution and various algorithms for its computation have been studied. These include the shooting algorithm (Fu, 1998), the least angle regression (LARS; Efron et al., 2004) and the coordinate descent algorithm (Friedman et al., 2007). Since the Lasso penalizes all the regression coefficients equally, it over-penalizes the important variables thereby resulting in biased estimators. The lack of the oracle property (Fan & Li, 2001) of Lasso prompted the development of the adaptive Lasso (Zou, 2006) with this property. The oracle property implies the method is consistent in variable selection, unbiased and asymptotically normal. The estimator is defined as

$$\hat{\boldsymbol{\theta}}_{\text{alasso}}(\lambda) = \underset{\theta}{\text{argmin}}\left\{-l(\boldsymbol{\theta}) + \lambda \sum_{d=1}^{p+q} w_d|\theta_d|\right\}, \tag{4.2}$$

where $\mathbf{w} = (\mathbf{w_1}, \ldots, \mathbf{w_{p+q}})$ is data-driven adaptive weight, which is often given as $\mathbf{w} = \mathbf{1}/|\hat{\boldsymbol{\theta}}|$ and $\hat{\boldsymbol{\theta}}$ is any consistent estimator of $\boldsymbol{\theta}$. We take this to be the maximum likelihood estimator, $\hat{\boldsymbol{\theta}}_{\text{ml}}$.

4.2 Least squares approximation

It is not straightforward to optimize the penalized log-likelihood function in equation (4.2). To simplify the optimization problem, we approximate $l(\boldsymbol{\theta})$ by the least squares approximation (LSA) method. Consider the second-order Taylor expansion of $l(\boldsymbol{\theta})$ at $\hat{\boldsymbol{\theta}}_{\text{ml}}$,

$$l(\boldsymbol{\theta}) \approx l(\hat{\boldsymbol{\theta}}_{\text{ml}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\text{ml}})^T l'(\hat{\boldsymbol{\theta}}_{\text{ml}}) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\text{ml}})^T l''(\hat{\boldsymbol{\theta}}_{\text{ml}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\text{ml}}), \tag{4.3}$$

where $l'(\cdot)$ and $l''(\cdot)$ are the first- and second-order derivatives of the log-likelihood function. Since $l(\hat{\boldsymbol{\theta}}_{\text{ml}})$ is a constant, $l'(\hat{\boldsymbol{\theta}}_{\text{ml}}) = 0$, and $l''(\hat{\boldsymbol{\theta}}_{\text{ml}}) = \hat{\Sigma}^{-1}$, where $\hat{\Sigma}$ is the estimated variance-covariance matrix of $\hat{\boldsymbol{\theta}}_{\text{ml}}$, we have

$$l(\boldsymbol{\theta}) \approx constant + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\text{ml}})^T \hat{\Sigma}^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\text{ml}}).$$

We fitted the models and obtain $\hat{\boldsymbol{\theta}}_{\text{ml}}$ and $\hat{\Sigma}$ by using the *GJRM* package in R (Marra & Radice, 2020). A pseudo data is created as

$$X^* = \hat{\Sigma}^{-1/2},\ Y^* = \hat{\Sigma}^{-1/2}\hat{\boldsymbol{\theta}}_{\text{ml}},$$

where X^* is a square matrix containing all the $(p+q+3)$ predictors along with the correlation parameter ρ and the intercepts, and Y^* is the corresponding pseudo response. Thus, equation (4.2) can be re-written as

$$\hat{\boldsymbol{\theta}}_{\text{alasso}}(\lambda) \approx \underset{\boldsymbol{\theta}}{\text{argmin}}\left\{\frac{1}{2}\left(Y^* - \boldsymbol{\theta}' X^*\right)^T\left(Y^* - \boldsymbol{\theta}' X^*\right) + \lambda\sum_{d=1}^{p+q} w_d(\left|\theta_d\right|)\right\},$$

which we optimized using the coordinate descent algorithm (see Friedman et al., 2010; Simon et al., 2011; Ogundimu, 2021).

4.3 Selection of tuning parameter

The optimal tuning parameter, λ can be estimated by using AIC (Akaike information criterion), BIC (Bayesian information criterion) and GCV (generalized cross-validation). It has been shown that the combination of the adaptive Lasso penalty and BIC-type tuning parameter selector results in LSA estimator that can be as efficient as the oracle estimator (Wang et al., 2007). Thus, we focus on the BIC criterion although the method is implemented for both AIC and GCV as well. The expression is given as

$$\text{BIC}(\lambda) = -2l(\hat{\boldsymbol{\theta}}) + df_\lambda \log(n),$$

where $0 \leq df_\lambda \leq (p+q)$ is the degree of freedom corresponding to the number of nonzero coefficients of $\hat{\boldsymbol{\theta}}$. The optimal value of λ is computed over a grid of candidate values of λ between $\lambda = 0$ and $\lambda = \lambda_{\text{max}}$, with step size of 0.1, where λ_{max} is the value of λ for which the entire vector of $\hat{\boldsymbol{\theta}}$ is zero. We allowed optimal $\lambda = 0$ for the unregularized solution.

4.4 Variance estimation

The variance of the nonzero component of $\hat{\boldsymbol{\theta}}$ (base on the optimal tuning parameter λ) can be derived using the local quadratic approximation (LQA) sandwich formula given in Fan & Li (2001). We suggest alternative formulation using block decomposition of the Hessian matrix $l''(\boldsymbol{\theta})$ (see Section A.2. in Supplementary Materials) and by generalization the Hessian matrix corresponding to equation (3.5). Let $\hat{\boldsymbol{\theta}}_1$ (with r elements, $r \leq s$) be non-vanishing component of $\hat{\boldsymbol{\theta}}$. Define

$A(\hat{\boldsymbol{\theta}}) = \text{diag}\{1/\hat{\boldsymbol{\theta}}_{11}, \ldots, 1/\hat{\boldsymbol{\theta}}_{1s}\}$. Let $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)$, where $\hat{\boldsymbol{\theta}}_1$ is as defined previously, and $\hat{\boldsymbol{\theta}}_2$ are the zero elements of $\hat{\boldsymbol{\theta}}$. Then,

$$M = \nabla^2 l(\hat{\boldsymbol{\theta}}) = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix},$$

where M_{11} corresponds to the first $r \times r$ submatrix of M. Further, let A_{11} be the first $r \times r$ submatrix of $A(\hat{\boldsymbol{\theta}})$. Define $E = M_{22} - M_{21}M_{11}^{-1}M_{12}$ and $\tilde{M}_{11} = M_{11} + \lambda A_{11}$. Then,

$$\widehat{\text{cov}}(\hat{\boldsymbol{\theta}}_1) = M_{11}^{-1} + \left(M_{11}^{-1} - \tilde{M}_{11}^{-1}\right) M_{12} E^{-1} M_{21} \left(M_{11}^{-1} - \tilde{M}_{11}^{-1}\right).$$

We have presented the variance estimation formula for the sake of completeness as the focus of the current work is on predictions. Further details on variance estimation for regularized sample selection model can be found in Ogundimu (2021).

5 Performance metrics and bootstrap validation

We describe the metrics for predictive accuracy and the bootstrap approach for model validation.

5.1 Metrics for predictive performance

In credit risk assessment, the misclassification of loan defaulters into non-defaulters will result in a loss for banks/creditors. Therefore, it is more important that the true defaulters are correctly classified. Here, we focus on model evaluation criteria for predictions in the context of regression analysis rather than classification. This is to ensure that the metrics for predictive performance are not threshold dependent and the users of the model can determine the appropriate threshold for classification. Unlike in previous studies, where area under the curve is the most common metric of prediction accuracy, we examined the performance of four other metrics based on model discrimination and calibration. The following performance metrics are used:

(1) *Area under the receiver operating curve* (AUROC): The c-index (Harrell et al., 1982) is the generalization of the AUROC, which is a measure of model performance that separates subjects with the event of interest from subjects without the event (discrimination). It calculates the proportion of pairs in which the predicted event probability is higher for the subject with the event of interest than that for the subject without the event. A model with no discriminatory ability has a value around 0.5 whereas a value close to 1 suggests excellent discrimination.

(2) *Area under the precision-recall curve* (AUPRC): Suppose True positive (TP) is defined as actual defaulters who are correctly predicted, False negative (FN) as actual defaulters who are predicted as non-defaulters, and False positive (FP) as actual non-defaulters predicted as defaulters. Then, $Recall = TP/(TP + FN)$ and $Precision = TP/(TP + FP)$. Thus, the precision-recall curve shows the relationship between precision and recall for every possible threshold value. The area under the curve is a single number summary of the information in the precision-recall (PR) curve. A key advantage of AUPRC is that it takes into consideration the prior probability of the outcome of interest, thereby reflecting the ability of the model to identify defaulters (often the minority class in credit scoring). Unlike the AUROC, its values range from 0 to 1. Its value approaches 0 as the prior probability of the outcome decreases (Davis & Goadrich, 2006). We computed this metric by using the *PRROC* package in R (Grau et al., 2015).

(3) *Brier score* (BS): It is a measure of agreement between the observed binary outcome (i.e., default versus non-default) and the predicted PD as shown below

$$\text{BS} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2,$$

where Y_i is the outcome and $\hat{Y}_i$ is the predicted PD. It is a proper scoring rule in that it is maximized when correct probabilities are used. A key advantage is that it captures both discrimination and calibration, and a low value of the metric is preferred.

(4) *Calibration Metrics*: We consider two metrics of calibration–*Expected Calibration Error* (ECE) and *Maximum Calibration Error* (MCE). The metrics are computed by sorting predicted probabilities and partitioning it into K fixed number of equal-frequency bins. The ECE calculates the average calibration error over the bins as

$$\text{ECE} = \sum_{i=1}^{K} P(i).|y_i - \hat{y}_i|,$$

where y_i is the true fraction of positive instances in bin i, $\hat{y}_i$ is the average of the probabilities for the instances in bin i, and $P(i)$ is the empirical probability (fraction) of all instances that fall into bin i, and the MCE calculates the maximum calibration error for the bins as

$$\text{MCE} = \max_{i=1,\cdots,K}|y_i - \hat{y}_i|.$$

The choices between $K = 10$ and $K = 100$ have been reported in the literature (Naeini et al., 2015; Wang et al., 2019). We chose $K = 10$ in this study. Like

Brier score, the lower the values of ECE and MCE, the better is the calibration of a model.

5.2 Bootstrap internal validation

Although we implemented penalized sample selection models to alleviate overfitting, some degree of optimism may persist, nonetheless. Harrell et al. (1996) presented a procedure for estimating optimism in predictive models. We extend this method to incorporate variable selection. Without loss of generality, consider a dataset $\mathcal{D} = \{\mathbf{x}_i, Y_i, S_i\}$, where $\mathbf{x}_i$, Y_i and S_i are as defined in Section 3.1, and performance metric P. A generic algorithm for the procedure is given in Algorithm 1.

Algorithm 1: Bootstrap validation with variable selection

Input: $\mathcal{D} = \{\mathbf{x}_i, Y_i, S_i\}$:
for $b = 1$ to B **do**
Take a bootstrap sample $\mathcal{D}_b$ from $\mathcal{D}$
fit model to $\mathcal{D}_b$ using regularized sample selection model
(grid search for optimal λ is done on each $\mathcal{D}_b$)
predict on the same $\mathcal{D}_b$ sample and compute predictive accuracy metric of interest, say $P^{\text{boot}}_{\text{boot}(b)}$
use the model to predict on $\mathcal{D}$ and compute predictive accuracy metric, $P^{\text{orig}}_{\text{boot}(b)}$
compute average optimism: Optimism $= \frac{1}{B}\sum_{b=1}^{B}\left(P^{\text{boot}}_{\text{boot}(b)} - P^{\text{orig}}_{\text{boot}(b)}\right)$
end for
fit model to $\mathcal{D}$ using regularized sample selection model
use the model to predict on $\mathcal{D}$ and compute apparent performance: $P^{\text{orig}}_{\text{orig}}$
Output: Optimism corrected metric P is $P^{\text{orig}}_{\text{orig}} -$ Optimism.

The optimism corrected metric P is the metric that has been corrected for overfitting. It is noteworthy that optimal lambda value is selected for each of the b bootstrap sample.

6 Numerical studies

In this section, we use simulation and a real data to evaluate the utility of the proposed estimators in reject inference.

6.1 Simulation study

We generated $Y_i^\star = \boldsymbol{\beta}^T \mathbf{x}_i + \varepsilon_{1i}$ and $S_i^\star = \boldsymbol{\gamma}^T \mathbf{w}_i + \varepsilon_{2i}$ as follows:

$$\boldsymbol{\beta}^T = (-2.78, 0.2, 0.2, 0.2, 0, 0, 0, 0, 0, 0.7, 0.7, 0.7)$$
$$\boldsymbol{\gamma}^T = (1.90, 0.2, 0.2, 0.2, 0, 0, 0, 0, 0, 0.7, 0.7, 0.7, 1).$$

The intercepts of the outcome equation, $\beta_0 = -2.78$ and the selection equation, $\gamma_0 = 1.90$ are chosen such that the required event rate and missing data is about 10% and 22% respectively. The 10% event rate is typical of datasets for modelling PD (Ogundimu, 2019). The simulation design ensures that there is one predictor in the selection equation that is not in the outcome equation (exclusion restriction–although this is not essential as demonstrated in Ogundimu, 2021). The covariates $\mathbf{x}_i$ and $\mathbf{w}_i$ are independent of the error terms $\varepsilon_i^T = (\varepsilon_{1i}, \varepsilon_{2i})$. We generated the underlying error distribution in two ways:

i BPSSM: the errors are generated with mean zero and correlation matrix $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, where $\rho = \{0, 0.2, 0.5\}$ ($\rho = 0$ corresponds to ignorable selection process)

ii CBSSM: the errors are generated from AMH copula with association measure $\theta_{\text{AMH}} = \{0, 0.498, 1\}$. Note that these values are equivalent to the values of ρ in the distribution of error terms under BPSSM. Specifically,

$$\tau = 0 \Rightarrow \rho = 0 \Rightarrow \theta_{\text{AMH}} = 0$$
$$\tau = 0.128 \Rightarrow \rho = 0.2 \Rightarrow \theta_{\text{AMH}} = 0.498$$
$$\tau = 0.333 \Rightarrow \rho = 0.5 \Rightarrow \theta_{\text{AMH}} = 1,$$

where τ is the Kendall's tau.

The covariates $x_1, \ldots, x_8$ are generated such that their distribution are marginally standard normal with pairwise correlations $\text{corr}(x_j, x_k) = \varrho^{|j-k|}$. We take $\varrho = 0.5$ to allow for moderate correlation between the covariates. The binary versions of $Y_i^\star$ and $S_i^\star$, denoted as Y_i and S_i respectively, are generated as in Section 3.1. Y_i is observed if $S_i = 1$ and missing otherwise. A sample of $n = 1000$ is used with 200 replications. We have combined weak (β = 0.2) and moderate ($\beta = 0.7$) covariates effects in this design.

We evaluated three classes of models: Bivariate probit model with sample selection (BPSSM), Copula bivariate sample selection model (CBSSM) and accept-only probit model (PROBIT). That is, adaptive Lasso penalized BPSSM (BPSSM_ALasso), Lasso penalized BPSSM (BPSSM_Lasso) and BPSSM with variables selected using p-value at 5% level of significance (BPSSM_P-value). For the copula-based model, we have adaptive Lasso penalized CBSSM (CBSSM_ALasso) and Lasso penalized CBSSM

Table 2 Results on the covariate selection based on selection model and corresponding complete case analyses–Bivariate normal data generation

Method	Sensitivity	Specificity	Sensitivity	Specificity
	Outcome equation		Selection equation	
		$\rho = 0$		
BPSSM_P-value	0.738	0.950	0.846	0.951
BPSSM_Lasso	0.955	0.689	0.990	0.606
BPSSM_ALasso	0.830	0.913	0.909	0.925
CBSSM_Lasso	0.963	0.698	0.989	0.625
CBSSM_ALasso	0.828	0.934	0.904	0.936
PROBIT_P-value	0.742	0.948	–	–
PROBIT_Lasso	0.967	0.700	–	–
PROBIT_ALasso	0.828	0.941	–	–
		$\rho = 0.2$		
BPSSM_P-value	0.713	0.955	0.844	0.946
BPSSM_Lasso	0.960	0.693	0.994	0.599
BPSSM_ALasso	0.812	0.928	0.904	0.932
CBSSM_Lasso	0.961	0.719	0.995	0.622
CBSSM_ALasso	0.810	0.941	0.905	0.938
PROBIT_P-value	0.726	0.955	–	–
PROBIT_Lasso	0.960	0.734	–	–
PROBIT_ALasso	0.795	0.950	–	–
		$\rho = 0.5$		
BPSSM_P-value	0.720	0.952	0.849	0.953
BPSSM_Lasso	0.962	0.701	0.995	0.628
BPSSM_ALasso	0.817	0.920	0.911	0.926
CBSSM_Lasso	0.966	0.716	0.994	0.634
CBSSM_ALasso	0.819	0.942	0.908	0.938
PROBIT_P-value	0.727	0.954	–	–
PROBIT_Lasso	0.963	0.729	–	–
PROBIT_ALasso	0.810	0.950	–	–

(CBSSM_Lasso), while the accept-only model includes probit model with adaptive Lasso (PROBIT_ALasso), probit model with Lasso (PROBIT_Lasso) and probit model with p-value (PROBIT_P-value).

We evaluated the performance of the methods using sensitivity (mean of proportion of nonzero coefficients that were correctly identified) and specificity (mean of proportion of zero coefficients that were correctly identified). The predictive accuracy of the model is evaluated using bootstrap method as described in Section 5. For each bootstrap sample, optimal λ is computed over a grid of candidate values of λ as described in Section 4.3 to provide a model having predictors and coefficients based on that penalty.

In Table 2, we present the results of the sensitivity and specificity of the methods for variable selection. Lasso methods have higher sensitivity but lower specificity than the other methods. This observation is not surprising since Lasso estimator lacks oracle property (Zou, 2006) and it is generally known to include true covari-

Table 3 Simulation results for the number of times each covariate is selected (out of 200) with both weak and moderate covariate effects (Outcome equation)–Bivariate normal data generation

Method	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
				$\rho = 0$							
BPSSM_P-value	109	84	93	8	4	13	13	12	200	200	200
BPSSM_Lasso	182	184	180	69	53	64	64	61	200	200	200
BPSSM_ALasso	140	126	130	20	9	23	16	19	200	200	200
CBSSM_Lasso	186	187	182	65	53	66	62	56	200	200	200
CBSSM_ALasso	138	125	130	12	7	17	15	15	200	200	200
Probit_P-value	110	85	95	9	7	11	14	11	200	200	200
Probit_Lasso	188	188	183	67	54	64	58	57	200	200	200
Probit_ALasso	138	127	129	14	6	16	10	13	200	200	200
				$\rho = 0.2$							
BPSSM_P-value	92	88	76	8	12	7	13	5	200	200	200
BPSSM_Lasso	186	190	176	62	58	65	61	61	200	200	200
BPSSM_ALasso	126	135	113	19	14	10	19	10	200	200	200
CBSSM_Lasso	186	191	176	50	54	60	59	58	200	200	200
CBSSM_ALasso	126	134	112	13	13	11	13	9	200	200	200
Probit_P-value	97	93	81	9	12	6	13	5	200	200	200
Probit_Lasso	187	190	175	52	49	55	56	54	200	200	200
Probit_ALasso	118	128	108	12	11	6	15	6	200	200	200
				$\rho = 0.5$							
BPSSM_P-value	107	81	76	9	10	7	14	8	200	200	200
BPSSM_Lasso	186	193	175	59	54	59	68	59	200	200	200
BPSSM_ALasso	132	136	112	19	17	11	19	14	200	200	200
CBSSM_Lasso	188	194	177	54	54	56	64	56	200	200	200
CBSSM_ALasso	131	138	114	10	15	7	16	10	200	200	200
Probit_P-value	109	84	79	8	10	8	12	8	200	200	200
Probit_Lasso	188	192	176	55	50	52	58	56	200	200	200
Probit_ALasso	129	135	108	9	11	6	15	9	200	200	200

ates, but also irrelevant covariates (Meinshausen & Bühlmann, 2006). Although the data was generated based on a bivariate normal distribution, the performance of CBSSM methods is superior to the corresponding BPSSM methods in terms of specificity. There is no clear distinction among the methods in terms of sensitivity. The adaptive Lasso methods have slightly better overall performance on the combined effects of sensitivity and specificity. To see the impact of weak and moderate covariate effects on the methods, we present the frequency with which the variables are selected in 200 replications in Table 3. The unregularized methods (BPSSM_P-value and Probit_P-value) selected true covariates less often when the effect is weak, but it is slightly better in selecting fewest covariates with true zero coefficients. CBSSM_Lasso is slightly better than BPSSM_Lasso across the correlation values. There is no clear advantage of sample selection models over complete case analyses in Tables 2 and 3.

Table 4 shows the result of quantifying optimistic predictions in the models. Regularized methods are expected to exhibit smaller optimism as the regularization

Table 4 Results of the optimism corrected model performance–Bivariate normal data generation

Method	AUROC	AUPRC	Brier	ECE	MCE
			$\rho = 0$		
BPSSM_P-value	0.929	0.650	0.057	0.014	0.057
BPSSM_Lasso	0.931	0.658	0.057	0.020	0.081
BPSSM_ALasso	0.930	0.655	0.056	0.015	0.059
CBSSM_Lasso	0.931	0.659	0.056	0.020	0.079
CBSSM_ALasso	0.931	0.655	0.056	0.015	0.058
Probit_P-value	0.929	0.650	0.057	0.014	0.058
Probit_Lasso	0.932	0.659	0.056	0.018	0.070
Probit_ALasso	0.931	0.656	0.056	0.015	0.056
			$\rho = 0.2$		
BPSSM_P-value	0.928	0.646	0.057	0.014	0.055
BPSSM_Lasso	0.931	0.655	0.056	0.019	0.078
BPSSM_ALasso	0.930	0.655	0.056	0.015	0.056
CBSSM_Lasso	0.931	0.656	0.056	0.019	0.076
CBSSM_ALasso	0.930	0.653	0.056	0.015	0.054
Probit_P-value	0.928	0.648	0.057	0.014	0.057
Probit_Lasso	0.931	0.657	0.056	0.018	0.073
Probit_ALasso	0.930	0.653	0.056	0.014	0.054
			$\rho = 0.5$		
BPSSM_P-value	0.928	0.647	0.057	0.013	0.052
BPSSM_Lasso	0.931	0.656	0.056	0.019	0.076
BPSSM_ALasso	0.930	0.653	0.056	0.015	0.055
CBSSM_Lasso	0.931	0.657	0.056	0.019	0.072
CBSSM_ALasso	0.930	0.653	0.056	0.014	0.055
Probit_P-value	0.928	0.648	0.057	0.014	0.054
Probit_Lasso	0.931	0.657	0.056	0.018	0.068
Probit_ALasso	0.930	0.653	0.056	0.014	0.053

is meant to alleviate the problem of overfitting. The results indicate that the use of sample selection models in reject inference problem, whether regularized or not, does not improve the accuracy of complete case analysis. Lasso-based methods are slightly better than the other methods in terms of the metrics for discrimination (AUROC and AUPRC). This is counterbalanced by its performance on the two metrics for calibration (ECE and MCE), where calibration results are consistently poorer than the other methods. This may be due to the inclusion of unimportant variables in Lasso methods.

Table A2 in the Supplementary Materials is equivalent to the results in Table 2 but with the data generated from AMH copula. Overall, CBSSM performance is slightly better than BPSSM (due to its performance on specificity). In general, complete case analyses are slightly better in terms of specificity for the outcome model whereas the regularized CBSSM sample selection models are better in terms of sensitivity. CBSSM_ALasso is better than BPSSM_ALasso (see Table A3 in Supplementary Materials). The results also show that there is slight benefit of the sample selection models

Table 5 Variable selection for the American Express credit card data (Default Models)

	Sample selection (BPSSM)			Complete cases (Probit)		
	P-value	Lasso	Adaptive Lasso	*P*-value	Lasso	Adaptive Lasso
(Intercept)	−1.056	−1.076	−1.070	−0.905	−0.934	−1.028
AGE	−0.006	−0.006	−0.005	−0.005	−0.005	–
ACADMOS	–	0.000	0.000	–	0.000	–
ADEPCNT	–	0.038	0.043	0.055	0.053	0.035
AEMPMOS	0.001	0.001	0.001	0.001	0.001	0.000
MAJORDRG	–	−0.019	−0.008	0.114	0.101	0.065
MINORDRG	0.079	0.073	0.074	0.082	0.079	0.055
OWNRENT	–	–	–	–	−0.008	–
APADMOS	–	0.000	0.000	–	0.000	–
AMAMIND	–	−0.081	−0.107	–	−0.094	–
INCOME	−0.000	−0.000	−0.000	−0.000	−0.000	−0.000
SELFEMPL	–	–	–	–	–	–
TRADACCT	–	0.006	0.005	–	−0.002	–
INCPER	–	−0.000	–	–	−0.000	–
EXP_INC	−0.382	−0.164	−0.352	−0.361	−0.110	–
CPTOPNB	0.013	0.012	0.012	0.016	0.015	0.005
CPTOPNG	−0.113	−0.112	−0.111	−0.122	−0.121	−0.104
CPT30C	0.246	0.217	0.244	0.297	0.261	0.264
CPTF30	0.065	0.065	0.065	0.086	0.087	0.087
CPTAVRV	–	0.003	0.002	–	0.001	–
CBURDEN	0.004	0.003	0.004	0.004	0.004	0.004

over the corresponding complete case analyses. However, the use of sample selection model does not translate to improved predictions (see Table A4 in Supplementary Materials).

6.2 Data analysis

We used the American Express credit card dataset that was described in Section 2 to evaluate the performance of our models. Table 5 gives the variables selected and parameter estimates of the bivariate probit model with sample selection (BPSSM) and classical probit model (PROBIT) for the default equation. The performance of BPSSM_Lasso and BPSSM_ALasso are similar in terms of the variables that are associated with PD except for the variable INCPER (income per family member), which is retained in the former but not the latter. BPSSM_P-value removed 10 variables from the default model, BPSSM_ALasso removed three variables and BPSSM_Lasso removed two variables. The variables OWNRENT and SELFEMPL are the two variables that are removed from the three models. The models under the complete case analyses (default model only) are based on variable selection using probit regression. Unlike in BPSSM models, PROBIT_ALasso shrinks more parameters (10 variables) to zero than PROBIT_P-value (eight variables). PROBIT_ALasso set AGE and EXP_INC to zero whereas these variables are associated with PD in PROBIT_P-value method. The only variable set to zero by PROBIT_Lasso is SELFEMPL.

Table 6 Variable selection for the American Express credit card data based on BPSSM (Cardholder Models)

	P-value	Lasso	Adaptive Lasso
(Intercept)	2.474	0.006	-0.009
AGE	–	−0.001	–
ACADMOS	0.001	0.001	0.001
ADEPCNT	−0.072	−0.070	−0.075
AEMPMOS	–	−0.000	−0.000
MAJORDRG	−0.777	−0.771	−0.776
MINORDRG	–	−0.037	−0.033
OWNRENT	–	–	–
APADMOS	–	0.000	0.000
AMAMIND	0.184	0.148	0.176
INCOME	0.000	0.000	0.000
SELFEMPL	−0.364	−0.331	−0.361
TRADACCT	0.110	0.109	0.109
INCPER	–	0.000	–
CPTOPNB	−0.022	−0.021	−0.021
CPTOPNG	0.032	0.032	0.030
CPT30C	−0.275	−0.256	−0.275
CPTF30	−0.086	−0.087	−0.087
CPTAVRV	0.008	0.008	0.007
CBURDEN	–	−0.001	−0.001
BANKSAV	–	−0.473	−0.508
BANKCH	–	–	–
BANKBOTH	–	0.483	0.488
CREDMAJR	0.295	0.285	0.295
ACBINQ	−0.180	−0.179	−0.180
ρ	0.530	0.536	0.519

Note: ρ: Correlation between disturbances.

The comparison across Table 5 of BPSSM and PROBIT methods show that INCOME, MINORDRG, AEMPMOS, CBURDEN, CPTOPNB, CPTOPNG, CPT30C and CPTF30 are important predictors of PD. The lower the income the more likely for an applicant to default while the higher the credit burden the more likely for the applicant to default. A striking observation is the setting of MAJORDRG to zero in BPSSM_P-value model whereas the variable is retained in other models across Table 5. However, all the methods (both under BPSSM and PROBIT) show that MINORDRG is associated with PD.

Table 6 shows the results of variable selection in the selection equation of BPSSM models. The performance of BPSSM_Lasso and BPSSM_ALasso are again similar in terms of the variables that are set to zero except for two variables–AGE and INCPER. Variables OWNRENT and BANKCH are not predictive of selection into the sample.

We also fitted the copula model (CBSSM) to the data. Table 7 shows the comparison of the models with BPMSS for the default model. The performance of Lasso methods is similar but coefficients from BPMSS_Lasso are shrunk more towards zero than CBSSM_Lasso. CBSSM_ALasso set 10 variables to zero whereas BPSSM_ALasso

Table 7 Variable selection for the American Express credit card data comparing CBSSM and BPSSM (Default Models)

	Sample selection (CBSSM)		Sample selection (BPSSM)	
	Lasso	Adaptive Lasso	Lasso	Adaptive Lasso
(Intercept)	−1.084	−1.162	−1.076	−1.070
AGE	−0.006	–	–0.006	−0.005
ACADMOS	0.000	–	0.000	0.000
ADEPCNT	0.038	0.024	0.038	0.043
AEMPMOS	0.001	0.000	0.001	0.001
MAJORDRG	−0.026	–	−0.019	−0.008
MINORDRG	0.074	0.056	0.073	0.074
OWNRENT	–	–	–	–
APADMOS	0.000	–	0.000	−0.000
AMAMIND	−0.083	–	−0.081	−0.107
INCOME	−0.000	−0.000	−0.000	−0.000
SELFEMPL	–	–	–	–
TRADACCT	0.006	–	0.006	0.005
INCPER	−0.000	–	−0.000	–
EXP_INC	−0.204	−0.151	−0.164	−0.352
CPTOPNB	0.013	0.007	0.012	0.012
CPTOPNG	−0.112	−0.095	−0.112	−0.111
CPT30C	0.227	0.222	0.217	0.244
CPTF30	0.066	0.059	0.065	0.065
CPTAVRV	0.002	–	0.003	0.002
CBURDEN	0.004	0.004	0.003	0.004

set only three variables to zero. Again, Lasso models for the cardholder equation are similar (Table 8).

Tables 9 and 10 show the predictive performance of the methods. There are still some amounts of optimism in the regularized methods. Interestingly, sample selection models are generally superior to complete case analyses in terms of discrimination (except for BPSSM_P-value)- a result that is not definitive from the simulation study.

7 Discussion

In this article, we introduced a variable selection technique based on lasso-type penalty for bivariate binary sample selection model. We also proposed a bootstrap internal validation method for this model. The sample selection models are analysed alongside complete case analyses (accept-only models). The simulation setting was structured to mimic typical rate of event and degree of missing data in practical data (10% event rate and 22% missing information). The results indicated that the proposed regularized sample selection model is suitable for variable selection in credit scoring research. We also concluded that the regularized results based on adaptive Lasso have better combined effects on sensitivity and specificity than the use of *p*-value, which is threshold dependent. This was the case in both the sample selection

Table 8 Variable selection for the American Express credit card data comparing CBSSM and BPSSM (Cardholder Models)

	Sample selection (CBSSM)		Sample selection (BPSSM)	
	Lasso	Adaptive Lasso	Lasso	Adaptive Lasso
(Intercept)	0.023	0.499	0.006	−0.009
AGE	−0.001	–	−0.001	–
ACADMOS	0.001	0.001	0.001	0.001
ADEPCNT	−0.071	−0.083	−0.070	−0.075
AEMPMOS	−0.000	–	−0.000	−0.000
MAJORDRG	−0.771	−0.773	−0.771	−0.776
MINORDRG	−0.038	–	−0.037	−0.033
OWNRENT	–	–	–	–
APADMOS	0.000	–	0.000	0.000
AMAMIND	0.153	0.107	0.148	0.176
INCOME	0.000	0.000	0.000	0.000
SELFEMPL	−0.336	−0.318	−0.331	−0.361
TRADACCT	0.109	0.107	0.109	0.109
INCPER	0.000	–	0.000	–
CPTOPNB	−0.021	−0.013	−0.021	−0.021
CPTOPNG	0.031	0.013	0.032	0.030
CPT30C	−0.255	−0.261	−0.256	−0.275
CPTF30	−0.085	−0.087	−0.087	−0.087
CPTAVRV	0.009	0.005	0.008	0.007
CBURDEN	−0.001	−0.000	−0.001	−0.001
BANKSAV	−0.478	−0.972	−0.473	−0.508
BANKCH	–	−0.478	–	–
BANKBOTH	0.484	–	0.483	0.488
CREDMAJR	0.285	0.288	0.285	0.295
ACBINQ	−0.179	−0.180	−0.179	−0.180
$\theta^{\star}$	0.919	0.918	0.536	0.519

Note: θ^{*}: Association measure between disturbances.

Table 9 Apparent model performance in predicting default probability in American Express credit card data

	AUROC	AUPRC	Brier	ECE	MCE
*BPSSM	0743	0.226	0.080	0.014	0.041
BPSSM_P-value	0.725	0.216	0.083	0.031	0.109
BPSSM_Lasso	0.743	0.226	0.080	0.015	0.040
BPSSM_ALasso	0.744	0.227	0.080	0.013	0.036
CBSSM_Lasso	0.743	0.226	0.080	0.013	0.042
CBSSM_ALasso	0.740	0.223	0.081	0.016	0.049
**Probit	0.734	0.218	0.081	0.012	0.037
Probit_P-value	0.733	0.218	0.081	0.013	0.036
Probit_Lasso	0.734	0.217	0.081	0.013	0.040
Probit_ALasso	0.731	0.215	0.081	0.015	0.055

Notes: *Sample selection model without variable selection.
**Probit model without variable selection.

Table 10 Optimism corrected model performance in predicting default probability in American Express credit card data

	AUROC	AUPRC	Brier	ECE	MCE
*BPSSM	0.737	0.218	0.081	0.012	0.035
BPSSM_P-value	0.720	0.208	0.083	0.030	0.105
BPSSM_Lasso	0.737	0.217	0.081	0.013	0.034
BPSSM_ALasso	0.738	0.219	0.081	0.012	0.030
CBSSM_Lasso	0.737	0.218	0.081	0.011	0.036
CBSSM_ALasso	0.734	0.215	0.081	0.014	0.043
**Probit	0.728	0.211	0.081	0.010	0.030
Probit_P-value	0.726	0.211	0.081	0.011	0.029
Probit_Lasso	0.728	0.210	0.081	0.011	0.032
Probit_ALasso	0.725	0.208	0.082	0.013	0.047

Notes: *Sample selection model without variable selection.
**Probit model without variable selection.

and the accept-only models. The simulation results for the internal validation of the prediction models did not provide definitive advantage of using sample selection model over the accept-only model. The cases where the metrics for discrimination are slightly better for the accept-only model were counterbalanced by the sample selection models doing relatively better on metrics for calibration. Overall, our results indicated that Lasso methods should be preferred for optimal predictions.

We have used the AMH copula function with Gaussian marginal distributions in this article. In application, we can incorporate the method of choosing a suitable copula and link function within the proposed framework. One way to do this is by optimizing optimism corrected predictive accuracy measure of interest (e.g., AUROC) over a suitable set of copulas and link functions with relatively small bootstrap samples (say, 20 to 30). What is needed in the current implementation is to add appropriate likelihood function for the copula and link function. The methods in this article are implemented in the R package *HeckmanSelect*, the package contains a simulated data (binHeckman) and the American Express credit card data (AmEX). It can be installed as follows:

```
devtools::install_github('EOgundimu300/HeckmanSelect').
```

We have used a simulation study with event rate and degree of selection that is similar to the American Express credit card dataset. It is unlikely that varying these factors will change our conclusions significantly. There are limitations of this study that deserved thorough attention. We have used a single penalty term with one tuning parameter for both the outcome and the selection equations, which is quite restrictive. Separate penalties can be used via approximation of the L_1 norm. Apart from Lasso and adaptive Lasso, the use of correlation-based penalty, like the one proposed in Tutz and Ulbricht (2009), can alleviate the problem of multicollinearity.

There are methods to incorporate more flexible covariate effect structures in sample selection models (e.g., splines and fractional polynomials). The method that we proposed can be readily extended to accommodate this flexibility by combining

LSA framework with group Lasso (Yuan & Lin, 2006). In this case, the flexible parameterization of covariates implies that the selection of one variable in a group will results in the selection of all other variables in the same group. Alternatively, the group bridge estimator (Huang et al., 2009) can be used, where simultaneous selection at both the group and within-group individual variable levels is possible.

Supplementary materials

Supplementary materials for this article are available at http://www.statmod.org/smij/archive.html.

Acknowledgements

The author thanks the editor, associate editor and the referees for their helpful comments which improved the article. The author is grateful to Professor William Greene for the helpful discussion on the conventional assumption of normality for the error terms in the classical binary sample selection model.

Declaration of conflicting interests

The author declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author received no financial support for the research, authorship and/or publication of this article.

References

Banasik J and Crook J (2007). Reject inference, augmentation, and sample selection. *European Journal of Operational Research*, **183**, 1582–1594.

Banasik J, Crook J, and Thomas L (2003). Sample selection bias in credit scoring models. *Journal of the Operational Research Society*, **54**, 822–832.

Chen GG and Astebro T (2012). Bound and collapse Bayesian reject inference for credit scoring. *Journal of the Operational Research Society*, **63**, 1374–1387.

Crook J and Banasik J (2004). Does reject inference really improve the performance of application scoring models? *Journal of Banking and Finance*, **28**, 857–874.

Davis J and Goadrich M (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on*

Machine Learning, ICML '06, pages 233–240. New York, NY: ACM. URL https://dl.acm.org/doi/10.1145/1143844.1143874 (last accessed 8 April 2022).

Dubin JA and Rivers D (1989). Selection bias in linear regression, logit and probit models. *Sociological Methods & Research*, **18**, 360–390.

Efron B, Hastie T, Johnstone I and Tibshirani R (2004). Least Angle Regression. *Annals of Statistics*, **32**, 407–499.

Fan J and Li R (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348–1360.

Friedman J, Hastie T and Tibshirani RJ (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**, 1–22.

Friedman J, Hastie T, Höfling H and Tibshirani R (2007). Pathwise coordinate optimization. *Annals of Statistics*, **1**, 302–332.

Fu WJ (1998). Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, **7**, 397–416.

Gomes M, Radice R, Brenes JC and Marra G (2019). Copula selection models for non-Gaussian outcomes that are missing not at random. *Statistics in Medicine*, **38**, 480–496.

Grau J, Grosse I and Keilwagen J (2015). PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics*, **31**, 2595–2597.

Greene WH (1998). Sample selection in credit-scoring models. *Japan and the World Economy*, **10**, 299–316.

Greene WH (2008). A statistical model for credit scoring. In *Advances in Credit Risk Modelling and Corporate Bankruptcy Prediction*, edited by S Jones and DA Hensher, pages 14–43. Cambridge: Cambridge University Press.

Hand DJ and Henley WE (1993). Can reject inference ever work? *IMA Journal of Management Mathematics*, **5**, 45–55.

Harrell FE, Califf RM, Pryor DB, Lee KL and Rosati RA (1982). Evaluating the yield of medical tests. *JAMA*, **247**, 2543–2546.

Harrell FE, Lee KL and Mark DB (1996). Tutorial in biostatistics, multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, **15**, 361–387.

Huang J, Ma S, Xie H and Zhang C-H (2009). A group bridge approach for variable selection. *Biometrika*, **96**, 339–355.

Kim Y and Sohn SY (2007). Technology scoring model considering rejected applicants and effect of reject inference. *Journal of the Operational Research Society*, **58**, 1341–1347.

Li Z, Tian Y, Li K, Zhou F and Yang W (2017). Reject inference in credit scoring using semi-supervised support vector machines. *Expert Systems With Applications*, **74**, 105–114.

Little RJA (1985). A note about models for selectivity bias. *Econometrica*, **53**, 1469–1474.

Marra G and Radice R (2020). Generalised Joint Regression Modelling [computer program]. *R package version* **0.2-3**. URL https://cran.r-project.org/web/packages/GJRM/GJRM.pdf (last accessed 8 April 2022).

Marra G, Radice R, Bärnighausen T, Wood SN and McGovern ME (2017b). A Simultaneous equation approach to estimating HIV prevalence with nonignorable missing responses. *Journal of the American Statistical Association*, **112**, 484–496.

Marshall A, Tang L and Milne A (2010). Variable reduction, sample selection bias and bank retail credit scoring. *Journal of Empirical Finance*, **17**, 501–512.

Meinshausen N and Bühlmann P (2006). High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, **34**, 1436–1462.

Naeini MP, Cooper GF and Hauskrecht M (2015). Obtaining well–calibrated probabilities using Bayesian binning. In Proc Conf AAAI Artif Intell 2015, pp. 2901–2907. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4410090/ (last accessed 8 April 2022).

Ogundimu EO (2019). Prediction of default probability by using statistical models for rare

events. *Journal of the Royal Statistical Society: Series A*, **182**, 1143–1162.

Ogundimu EO (2021). Regularization and variable selection in Heckman selection model. *Statistical Papers*, **63**, 421–439. URL https://doi.org/10.1007/s00362-021-01246-z (last accessed 8 April 2022).

Puhani PA (2000). The Heckman correction for sample selection and its critique. *Journal of Economic Surveys*, **14**, 53–68.

Simon N, Friedman J, Hastie T and Tibshirani R (2011). Regularization Paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, **39**, 1–13.

Tibshirani R (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society*, **58**, 267–288.

Tutz G and Ulbricht J (2009). Penalized regression with correlation-based penalty. *Statistics and Computing*, **19**, 239–253.

Wang H and Leng C (2007). Unified LASSO Estimation by Least Squares Approximation. *Journal of the American Statistical Association*, **102**, 1039–1048.

Wang H, Li R and Tsai C-L (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, **94**, 553–568.

Wang Y, Li L and Dang C (2019). Calibrating classification probabilities with shape-restricted polynomial regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **41**, 1823–1827.

Wu I, D and Hand DJ (2007). Handling selection bias when choosing actions in retail credit applications. *European Journal of Operational Research*, **183**, 1560–1568.

Yuan M and Lin Y (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*, **68**, 49–67.

Zeng P, Wei Y, Zhao Y, Liu J, Liu L, Zhang R, Gou J, Huang S and Chen F (2014). Variable selection approach for zero-inflated count data via adaptive Lasso. *Journal of Applied Statistics*, **41**, 879–894.

Zou H (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, **101**, 1418–1429.