Communications for Statistical Applications and Methods
2024, Vol. 31, No. 4, 441–457

https://doi.org/10.29220/CSAM.2024.31.4.441
Print ISSN 2287-7843 / Online ISSN 2383-4757

Variable selection and prediction performance of penalized two-part regression with community-based crime data application

Seong-Tae Kim[a], Man Sik Park[1,b]

[a]Department of Mathematics & Statistics, NC A&T State University, USA;
[b]Department of Statistics, Sungshin Women's University, Korea

Abstract

Semicontinuous data are characterized by a mixture of a point probability mass at zero and a continuous distribution of positive values. This type of data is often modeled using a two-part model where the first part models the probability of dichotomous outcomes -zero or positive- and the second part models the distribution of positive values. Despite the two-part model's popularity, variable selection in this model has not been fully addressed, especially, in high dimensional data. The objective of this study is to investigate variable selection and prediction performance of penalized regression methods in two-part models. The performance of the selected techniques in the two-part model is evaluated via simulation studies. Our findings show that LASSO and ENET tend to select more predictors in the model than SCAD and MCP. Consequently, MCP and SCAD outperform LASSO and ENET for β-specificity, and LASSO and ENET perform better than MCP and SCAD with respect to the mean squared error. We find similar results when applying the penalized regression methods to the prediction of crime incidents using community-based data.

Keywords: two-part model, semicontinuous data, variable selection, penalized regression, LASSO, SCAD, MCP, crime data, prediction

1. Introduction

As a subclass of non-negative data, semicontinuous data comprises a continuous distribution of positive values and a probability mass function of a nonnegligible number of zero values. We encounter semicontinuous data widely in various applications. Health care expenditures and medical costs are well-known applications (Duan *et al.*, 1983; Liu, 2009; Mullahy, 1998; Smith *et al.*, 2014; Tu and Zhou, 1999), where a considerable portion of healthy people incur no health care expenditures, while clinic or hospital visitors spend a wide range of medical costs. Other applications include substance abuse (Brown *et al.*, 2005), alcohol consumption (Olsen and Schafer, 2001), and smoking behavior (Zhao *et al.*, 2016).

The presence of two heterogeneous distributions of semicontinuous data makes ordinary least squares (OLS) estimation biased and inefficient. The two-part model (TPM) has gained popularity as an alternative to OLS. The TPM separately models the binary response (either zero or positive) and the mean response given that it is positive (Duan *et al.*, 1983; Cragg, 1971). Recent TPM studies

Kim is partially supported by NSF Grants 1719498 and 2100729.
[1]Corresponding author: Department of Statistics, Sungshin Women's University, 2 Bomun-ro 34da-gil, Seongbuk-gu, Seoul 02844, Korea. E-mail: mansikpark@sungshin.ac.kr

Published 31 July 2024 / journal homepage: http://csam.or.kr

include deep learning-based feature importance (Zou *et al.*, 2023), TPM quantile regression (Merlo *et al.*, 2022), and multivariate TPM model (Frees *et al.*, 2013). Similar to the TPM, Tweedie models can model semicontinuous data using a single unified distribution, called a Tweedie distribution (Dunn and Smyth, 2005; Kokonendji *et al.*, 2021; Tweedie, 1984). Despite the widespread use of the TPM for semicontinuous data, researchers have paid relatively little attention to variable selection and prediction performance of the TPM. Variable selection of identifying a best subset of predictors is a backbone of predictive regression modeling, and it is increasingly important to deal with high dimensionality in the era of big data. Our study aims to explore the variable selection and prediction performance of the two-part model using selected regularized regression methods in high-dimensional data.

Penalized (or regularized) regressions are useful in high-dimensional data in which a few number of predictors may contribute to modeling the response. Penalized regressions simultaneously perform variable selection and parameter estimation. Popular penalized regression methods based on soft thresholdinging include the least angle shrinkage and selection operator (LASSO) (Tibshirani, 1996), elastic net (ENET) (Zou and Hastie, 2005), adaptive LASSO (Zou, 2006), group LASSO (Yuan and Lin, 2006), Dantzig selector (Candes and Tao, 2007), smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), and minimax concave penalty (MCP) (Zhang, 2010). Since recent variations of penalized regressions (Hao *et al.*, 2018; Fan and Lv, 2008) mainly incorporate LASSO, ENET, SCAD, and MCP, we focus on these methods for the variable selection and prediction of TPM. We will compare these methods to traditional variable selection methods such as Akaike information criterion (AIC) and Bayesian information criterion (BIC) in simulation and empirical studies. The selected penalized regressions are estimated using the coordinate descent algorithm (CDA) was introduced (Wu and Lange, 2008; Breheny and Huang, 2011). The CDA optimizes a single parameter with others held fixed, cycling until the solution path stabilizes.

The TPM offers several intriguing characteristics. First, the two models for binary outcome and continuous positive response utilize the same predictor variables. However, the significance of individual predictors may differ between the two response models. Second, modeling the binary response uses a larger number of observations than the continuous positive response. In other words, not only is the entire sample size important, but the proportion of the positive values is important to modeling TPM. The conditional linear model for the continuous positive response may suffer from a smaller sample size. Last, the marginal mean of the response is a function of the probability that the responses are positive and the marginal mean of the positive responses. These characteristics deserve thorough investigation while performing variable selection and prediction in the TPM.

The simulation study probes the effects of various statistical scenarios in controlled settings. Our empirical study considers the prediction of crime incidents. In recent years, there have been significant advances in predictive modeling of crime incidents because of its societal implications and importance. Crime prediction has employed a broad spectrum of machine learning algorithms and data sources. These prediction algorithms encompass nonparameteric regression, support vector machine, and deep neural network, and the crime data include spatiotemporal data, social media data, and community-based data (Kang and Kang, 2017 and references therein). Our predictive modeling incorporates penalized regression using community-based data. The regression approach has advantages of identification and interpretation of the predictors associated with crime incidents.

The remaining sections are as follows. Section 2 briefly introduces the two-part model and its mean squared error. Section 3 describes regularized regression-based variable method and estimation. Section 4 describes the design and result of simulation study using the methods. Section 5 implements empirical data analysis using the community-based crime data. Last, Section 6 discusses the findings

Table 1: Variable selection of parameter space P1 in the two-part model for two covariance structures

n	p	Method	**Independent covariance structure**			
			Logistic regression		Linear regression	
			β-sensitivity	β-specificity	β-sensitivity	β-specificity
500	20	AIC	1(0)	0.832(0.007)	0.838(0.016)	0.803(0.008)
		BIC	1(0)	0.984(0.002)	0.983(0.006)	0.979(0.002)
		LASSO	1(0)	0.774(0.021)	1(0)	0.590(0.014)
		ENET	1(0)	0.720(0.021)	1(0)	0.443(0.013)
		SCAD	1(0)	0.938(0.007)	1(0)	0.914(0.010)
		MCP	1(0)	0.964(0.006)	1(0)	0.957(0.007)
	100	LASSO	1(0)	0.930(0.007)	1(0)	0.828(0.006)
		ENET	1(0)	0.896(0.008)	1(0)	0.743(0.007)
		SCAD	1(0)	0.975(0.003)	1(0)	0.961(0.003)
		MCP	1(0)	0.989(0.002)	1(0)	0.982(0.002)
	1000	LASSO	1(0)	0.992(0.001)	1(0)	0.977(0.001)
		ENET	1(0)	0.990(0.001)	1(0)	0.968(0.001)
		SCAD	1(0)	0.994(0.000)	1(0)	0.990(0.001)
		MCP	1(0)	0.998(0.000)	1(0)	0.998(0.000)
1000	20	AIC	1(0)	0.834(0.006)	0.840(0.015)	0.812(0.008)
		BIC	1(0)	0.989(0.002)	0.993(0.003)	0.985(0.002)
		LASSO	1(0)	0.806(0.019)	1(0)	0.587(0.014)
		ENET	1(0)	0.732(0.020)	1(0)	0.428(0.013)
		SCAD	1(0)	0.952(0.007)	1(0)	0.949(0.008)
		MCP	1(0)	0.954(0.007)	1(0)	0.960(0.007)
	100	LASSO	1(0)	0.914(0.009)	1(0)	0.849(0.006)
		ENET	1(0)	0.888(0.010)	1(0)	0.766(0.006)
		SCAD	1(0)	0.981(0.003)	1(0)	0.976(0.003)
		MCP	1(0)	0.990(0.002)	1(0)	0.991(0.002)
	1000	LASSO	1(0)	0.993(0.001)	1(0)	0.983(0.001)
		ENET	1(0)	0.988(0.001)	1(0)	0.972(0.001)
		SCAD	1(0)	0.997(0.000)	1(0)	0.996(0.001)
		MCP	1(0)	0.999(0.000)	1(0)	0.999(0.000)
			AR(1) covariance structure			
500	20	AIC	1(0)	0.810(0.007)	0.817(0.016)	0.801(0.008)
		BIC	1(0)	0.987(0.002)	0.978(0.006)	0.979(0.003)
		LASSO	1(0)	0.814(0.017)	1(0)	0.669(0.013)
		ENET	1(0)	0.694(0.021)	1(0)	0.538(0.013)
		SCAD	1(0)	0.921(0.007)	1(0)	0.916(0.008)
		MCP	1(0)	0.970(0.005)	1(0)	0.953(0.007)
	100	LASSO	1(0)	0.938(0.008)	1(0)	0.868(0.005)
		ENET	1(0)	0.910(0.008)	1(0)	0.804(0.006)
		SCAD	1(0)	0.957(0.003)	1(0)	0.955(0.003)
		MCP	1(0)	0.984(0.002)	1(0)	0.985(0.002)
	1000	LASSO	1(0)	0.992(0.001)	1(0)	0.985(0.001)
		ENET	1(0)	0.992(0.001)	1(0)	0.975(0.001)
		SCAD	1(0)	0.992(0.000)	1(0)	0.991(0.001)
		MCP	1(0)	0.998(0.000)	1(0)	0.998(0.000)
1000	20	AIC	1(0)	0.820(0.007)	0.832(0.015)	0.807(0.008)
		BIC	1(0)	0.989(0.002)	0.988(0.005)	0.986(0.002)
		LASSO	1(0)	0.797(0.019)	1(0)	0.657(0.013)
		ENET	1(0)	0.672(0.020)	1(0)	0.523(0.013)
		SCAD	1(0)	0.945(0.007)	1(0)	0.943(0.009)
		MCP	1(0)	0.969(0.005)	1(0)	0.967(0.007)
	100	LASSO	1(0)	0.923(0.008)	1(0)	0.872(0.006)
		ENET	1(0)	0.896(0.008)	1(0)	0.811(0.006)
		SCAD	1(0)	0.970(0.003)	1(0)	0.975(0.003)
		MCP	1(0)	0.987(0.002)	1(0)	0.991(0.002)
	1000	LASSO	1(0)	0.995(0.001)	1(0)	0.984(0.001)
		ENET	1(0)	0.993(0.001)	1(0)	0.976(0.001)
		SCAD	1(0)	0.995(0.000)	1(0)	0.995(0.000)
		MCP	1(0)	0.998(0.000)	1(0)	0.999(0.000)

The table reports the simulation mean (standard error) based on 200 iterations.

and related issues from both simulation study and empirics, concludes the study, and suggests further research areas.

2. Two-part model

Let $\mathbf{y} = (y_1, \ldots, y_n)^T$ be an $n \times 1$ response vector, where $y_i, i = 1, \ldots, n$, are independently observed semicontinuous responses such that, for $0 < \pi_i < 1$,

$$\begin{cases} y_i > 0 & \text{with probability } \pi_i, \\ y_i = 0 & \text{with probability } 1 - \pi_i. \end{cases}$$

Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ be the $n \times p$ matrix of predictors, where $\mathbf{x}_j$ is a $n \times 1$ vector of the j^{th} predictor for $j = 1, \ldots, p$. The TPM requires two sets of parameter spaces, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^T$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$. Let $\mathcal{A}_1 = \{j : \alpha_j \neq 0\}$ with cardinality $|\mathcal{A}_1| = p_1$, and $\mathcal{A}_2 = \{j : \beta_j \neq 0\}$ with cardinality $|\mathcal{A}_2| = p_2$, where j is an element of the index set $\mathcal{A} = \{1, 2, \ldots, p\}$. Then $\boldsymbol{\alpha}_{\mathcal{A}_1}$ and $\boldsymbol{\beta}_{\mathcal{A}_2}$ are subvectors of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. Also, $\mathbf{X}_{\mathcal{A}_1}$ and $\mathbf{X}_{\mathcal{A}_2}$ be submatrices of $\mathbf{X}$. We also define two indicator functions associated with the response as follows:

$$I_0 = \begin{cases} 1 & \text{if } y_i = 0 \\ 0 & \text{if } y_i > 0 \end{cases} \quad \text{and} \quad I_1 = \begin{cases} 0 & \text{if } y_i = 0 \\ 1 & \text{if } y_i > 0. \end{cases}$$

A conventional two-part model of y_i is defined as

$$f(y_i) = (1 - \pi_i)^{I_0} \times \left[\pi_i h\left(y_i \mid y_i > 0, \mathbf{x}_i\right)\right]^{I_1}, \quad i = 1, 2, \ldots, n, \tag{2.1}$$

where $h(y_i|y_i > 0, \mathbf{x}_i)$ is any continuous density function for $y_i > 0$. In the TPM in (2.1), $\pi_i = P(Y_i > 0)$ can be modeled by a parametric binary probability model such as logit or probit using predictors, that is, $\pi_i = P(Y_i > 0|\mathbf{x}_i)$. In this article, the logit model is used as follows:

$$\text{logit}(\pi_i) := \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \mathbf{x}_i^T \boldsymbol{\alpha}. \tag{2.2}$$

The logistic regression in (2.2) is the first part of the TPM. Consequently, the positive values of y_i are modeled by the parameterization of $h(y_i|y_i > 0, \mathbf{x}_i)$ as follows:

$$h\left(y_i \mid y_i > 0, \mathbf{x}_i\right) = \ln(y_i) = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \tag{2.3}$$

where the error terms ε_i are independent and identically normally distributed with mean 0 and variance σ^2. The positive values of the response are often skewed right, so they are modeled through a log-transformation of Y. The log-normal assumption is widely applied, but it is often restrictive, which can be relaxed by the log-skew-normal assumption (Smith *et al.*, 2014). This linear model is the second part of the TPM. The TPM of (2.2) and (2.3) is sometimes called the Bernoulli-log-normal regression model (Neelon *et al.*, 2016). The two regression models of the TPM are independently fitted. We can see the independence of fitting the two models via their likelihood functions. The likelihood function of the two-part model for the random sample of n independent observations is

$$\begin{aligned} L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma) &= \prod_{i=1}^{n} f(y_i) = \left\{ \prod_{i=1}^{n} (1 - \pi_i)^{I_0} \times \left[\pi_i h(y_i \mid y_i > 0, \mathbf{x}_i)\right]^{I_1} \right\} \\ &= \left\{ \prod_{i=1}^{n} (1 - \pi_i)^{I_0} \cdot \pi_i^{I_1} \right\} \times \left\{ \prod_{i=1}^{n} h(y_i \mid y_i > 0, \mathbf{x}_i)^{I_1} \right\} \\ &= L(\boldsymbol{\alpha}) \times L(\boldsymbol{\beta}, \sigma). \end{aligned} \tag{2.4}$$

Table 2: Prediction performance of parameter space P1 for two covariance structures

n	p	Method	Mean squared error		Classification		
			Positive	All	Accuracy	Sensitivity	Specificity
Independent covariance structure							
500	20	AIC	23.08(0.21)	23.83(0.22)	0.874(0.001)	0.852(0.002)	0.889(0.001)
		BIC	22.68(0.21)	23.24(0.21)	0.877(0.001)	0.857(0.002)	0.892(0.001)
		LASSO	23.17(0.20)	24.12(0.19)	0.876(0.001)	0.865(0.002)	0.884(0.001)
		ENET	23.21(0.20)	24.16(0.19)	0.874(0.001)	0.871(0.002)	0.877(0.002)
		SCAD	24.04(0.19)	25.24(0.19)	0.878(0.001)	0.858(0.002)	0.893(0.001)
		MCP	25.34(0.18)	25.23(0.18)	0.879(0.001)	0.859(0.002)	0.892(0.001)
	100	LASSO	22.91(0.2)	23.62(0.20)	0.862(0.001)	0.861(0.002)	0.864(0.002)
		ENET	22.87(0.21)	23.56(0.20)	0.859(0.001)	0.876(0.002)	0.851(0.002)
		SCAD	23.77(0.18)	24.83(0.17)	0.867(0.001)	0.845(0.002)	0.881(0.001)
		MCP	24.84(0.16)	24.85(0.17)	0.867(0.001)	0.845(0.002)	0.881(0.001)
	1000	LASSO	23.55(0.22)	24.16(0.21)	0.860(0.001)	0.865(0.002)	0.859(0.002)
		ENET	23.52(0.22)	24.08(0.22)	0.852(0.001)	0.885(0.002)	0.837(0.002)
		SCAD	24.90(0.19)	26.04(0.18)	0.867(0.001)	0.846(0.002)	0.882(0.001)
		MCP	26.14(0.17)	25.96(0.18)	0.868(0.001)	0.846(0.002)	0.882(0.001)
1000	20	AIC	25.62(0.26)	26.18(0.25)	0.875(0.001)	0.853(0.001)	0.891(0.001)
		BIC	25.58(0.26)	26.11(0.26)	0.878(0.001)	0.854(0.001)	0.894(0.001)
		LASSO	27.19(0.23)	28.40(0.23)	0.877(0.001)	0.863(0.001)	0.886(0.001)
		ENET	27.08(0.23)	28.30(0.23)	0.876(0.001)	0.867(0.001)	0.882(0.001)
		SCAD	28.39(0.24)	29.76(0.24)	0.878(0.001)	0.854(0.001)	0.894(0.001)
		MCP	29.99(0.24)	29.76(0.24)	0.878(0.001)	0.854(0.001)	0.894(0.001)
	100	LASSO	25.70(0.20)	26.59(0.19)	0.868(0.001)	0.858(0.001)	0.874(0.001)
		ENET	25.61(0.21)	26.44(0.20)	0.867(0.001)	0.868(0.002)	0.867(0.001)
		SCAD	26.96(0.17)	28.21(0.16)	0.870(0.001)	0.850(0.001)	0.883(0.001)
		MCP	28.50(0.15)	28.22(0.16)	0.870(0.001)	0.850(0.001)	0.883(0.001)
	1000	LASSO	25.03(0.22)	25.69(0.21)	0.873(0.001)	0.871(0.002)	0.874(0.001)
		ENET	24.97(0.22)	25.54(0.22)	0.868(0.001)	0.885(0.002)	0.860(0.001)
		SCAD	26.56(0.17)	27.72(0.16)	0.875(0.001)	0.856(0.002)	0.888(0.001)
		MCP	27.85(0.15)	27.78(0.16)	0.875(0.001)	0.856(0.002)	0.888(0.001)
AR(1) Covariance structure							
500	20	AIC	26.66(0.23)	27.54(0.24)	0.884(0.001)	0.866(0.002)	0.898(0.001)
		BIC	26.02(0.22)	26.62(0.22)	0.889(0.001)	0.872(0.002)	0.902(0.001)
		LASSO	26.66(0.19)	27.72(0.18)	0.887(0.001)	0.876(0.002)	0.896(0.001)
		ENET	26.67(0.19)	27.72(0.18)	0.886(0.001)	0.876(0.002)	0.893(0.002)
		SCAD	27.42(0.18)	28.64(0.17)	0.889(0.001)	0.873(0.002)	0.902(0.001)
		MCP	28.74(0.16)	28.62(0.17)	0.889(0.001)	0.873(0.002)	0.902(0.001)
	100	LASSO	26.98(0.21)	27.82(0.20)	0.876(0.001)	0.875(0.002)	0.877(0.002)
		ENET	26.86(0.22)	27.67(0.20)	0.873(0.001)	0.885(0.002)	0.867(0.002)
		SCAD	28.09(0.19)	29.30(0.18)	0.880(0.001)	0.867(0.002)	0.889(0.001)
		MCP	29.30(0.16)	29.31(0.18)	0.880(0.001)	0.867(0.002)	0.889(0.001)
	1000	LASSO	27.19(0.22)	27.88(0.21)	0.880(0.001)	0.876(0.002)	0.884(0.002)
		ENET	27.12(0.22)	27.74(0.22)	0.875(0.001)	0.888(0.002)	0.868(0.002)
		SCAD	28.97(0.17)	30.29(0.16)	0.880(0.001)	0.862(0.002)	0.894(0.001)
		MCP	30.21(0.15)	30.21(0.16)	0.882(0.001)	0.864(0.002)	0.896(0.001)
1000	20	AIC	28.73(0.22)	29.36(0.22)	0.886(0.001)	0.867(0.001)	0.900(0.001)
		BIC	28.64(0.22)	29.20(0.22)	0.888(0.001)	0.870(0.001)	0.902(0.001)
		LASSO	30.20(0.19)	31.43(0.19)	0.886(0.001)	0.873(0.001)	0.896(0.001)
		ENET	30.16(0.20)	31.36(0.19)	0.886(0.001)	0.875(0.001)	0.894(0.001)
		SCAD	31.16(0.19)	32.56(0.19)	0.888(0.001)	0.870(0.001)	0.902(0.001)
		MCP	32.61(0.17)	32.56(0.19)	0.888(0.001)	0.869(0.001)	0.902(0.001)
	100	LASSO	30.09(0.21)	31.24(0.20)	0.883(0.001)	0.871(0.001)	0.892(0.001)
		ENET	29.90(0.21)	30.98(0.20)	0.881(0.001)	0.877(0.001)	0.885(0.001)
		SCAD	31.77(0.18)	33.24(0.18)	0.884(0.001)	0.865(0.001)	0.898(0.001)
		MCP	33.43(0.17)	33.25(0.18)	0.884(0.001)	0.866(0.001)	0.898(0.001)
	1000	LASSO	29.50(0.26)	30.36(0.25)	0.888(0.001)	0.880(0.001)	0.894(0.001)
		ENET	29.32(0.26)	30.07(0.25)	0.885(0.001)	0.888(0.001)	0.884(0.001)
		SCAD	31.15(0.24)	32.48(0.23)	0.889(0.001)	0.874(0.001)	0.901(0.001)
		MCP	32.66(0.23)	32.51(0.23)	0.889(0.001)	0.874(0.001)	0.901(0.001)

The table reports the simulation mean (standard error) based on 200 iterations.

The likelihood function of the two-part model in (2.4) is decomposed into two multiplicative terms (Duan *et al.*, 1983; Min and Agresti, 2002). The first term solely depends on the parameters, $\boldsymbol{\alpha}$, in

(2.2), and the second term solely depends on the parameters, $\boldsymbol{\beta}$, as in (2.3). This multiplicity allows to separately achieve the maximum likelihood estimate (MLE) of $\boldsymbol{\alpha}$ and the MLE of $\boldsymbol{\beta}$ and σ for the TPM. The log-likelihood function of TPM, denoted $l(\cdot) = \log L(\cdot)$, is given as

$$l(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma) = l(\boldsymbol{\alpha}) + l(\boldsymbol{\beta}, \sigma). \tag{2.5}$$

The expected value of the response variable, Y, is defined as

$$\begin{aligned} E[Y \mid \mathbf{x}] &= E[Y \mid Y > 0, \mathbf{x}] \times P(Y > 0) + E[Y \mid Y = 0, \mathbf{x}] \times P(Y = 0) \\ &= E[Y \mid Y > 0, \mathbf{x}] \times P(Y > 0). \end{aligned} \tag{2.6}$$

Plugging (2.2) and the expected value of (2.3) into (2.6) yields

$$E[Y \mid \mathbf{x}] = \left\{1 + \exp\left(-\mathbf{x}^T \boldsymbol{\alpha}\right)\right\}^{-1} \mathbf{x}^T \boldsymbol{\beta}. \tag{2.7}$$

As mentioned in the introduction, the expected value in (2.7) is a function of the probability that $Y > 0$ and the expected value of Y for $y > 0$. Furthermore, the independence of the likelihood functions of two part model leads to the predicted value of Y for the TPM as follows:

$$\widehat{Y} = \left\{1 + \exp\left(-\mathbf{x}^T \widehat{\boldsymbol{\alpha}}\right)\right\}^{-1} \mathbf{x}^T \widehat{\boldsymbol{\beta}}, \tag{2.8}$$

where $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ can be estimated by various estimation methods, which are described in the following section. The mean squared error (MSE) is defined as

$$\text{MSE} = E\left[\left(Y - \widehat{Y}\right)^2 \mid Y > 0\right], \tag{2.9}$$

which will be used to evaluate prediction performance in the following sections.

3. Variable selection with regularized TPM

In this section, we will discuss how we select and estimate $\boldsymbol{\alpha}_{\mathcal{A}_1}$ and $\boldsymbol{\beta}_{\mathcal{A}_2}$, each of which are subvectors of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. Variable selection is a key analytical component in predictive regression analysis. Numerous variable selection techniques have been proposed, and this trend will remain steady in the data-centered era. Variable selection techniques include stepwise methods (forward, backward, and stepwise selection methods), pretesting method (univariate t-test), prediction-oriented criteria (minR^2, maxR^2, Mallow's C_p, and adjusted R^2), information criterion methods (AIC, BIC, and AIC for a small-sample correction (AIC$_c$)), and penalized regression methods (LASSO, Ridge, ENET, SCAD, and MCP). Limited space restricts us to cover selected methods that have been the most used in practice, which include four penalized regression methods (LASSO, ENET, SCAD, and MCP).

For the purpose of variable selection and prediction via penalized regression, we consider a penalized log-likelihood of TPM, denoted $l_p(\cdot)$, using (2.5) as follows:

$$l_p\left(\boldsymbol{\alpha}_{\mathcal{A}_1}, \boldsymbol{\beta}_{\mathcal{A}_2}, \sigma\right) = l\left(\boldsymbol{\alpha}_{\mathcal{A}_1}\right) - P_1\left(\boldsymbol{\alpha}_{\mathcal{A}_1}\right) + l\left(\boldsymbol{\beta}_{\mathcal{A}_2}, \sigma\right) - P_2\left(\boldsymbol{\beta}_{\mathcal{A}_2}\right), \tag{3.1}$$

where the nonnegative penalty functions are defined as

$$P_1\left(\boldsymbol{\alpha}_{\mathcal{A}_1}\right) = \sum_{j \in \mathcal{A}_1} P\left(\lambda, \gamma; \left|\alpha_j\right|\right) \quad \text{and} \quad P_2\left(\boldsymbol{\beta}_{\mathcal{A}_2}\right) = \sum_{j \in \mathcal{A}_2} P\left(\lambda, \gamma; \left|\beta_j\right|\right),$$

Table 3: Variable selection of parameter space P2 via 200 for two covariance structures

n	p	Method	Independent covariance structure			
			Logistic regression		Linear regression	
			β-sensitivity	β-specificity	β-sensitivity	β-specificity
500	20	AIC	0.944(0.004)	0.815(0.012)	0.823(0.008)	0.809(0.013)
		BIC	0.847(0.004)	0.986(0.004)	0.982(0.002)	0.984(0.004)
		LASSO	0.976(0.003)	0.426(0.024)	0.994(0.001)	0.129(0.012)
		ENET	0.981(0.003)	0.388(0.024)	0.994(0.001)	0.098(0.011)
		SCAD	0.959(0.004)	0.663(0.021)	0.965(0.003)	0.485(0.022)
		MCP	0.941(0.005)	0.758(0.020)	0.954(0.004)	0.549(0.026)
	100	LASSO	0.928(0.005)	0.801(0.011)	0.941(0.004)	0.579(0.007)
		ENET	0.930(0.005)	0.772(0.011)	0.946(0.003)	0.514(0.007)
		SCAD	0.906(0.004)	0.909(0.003)	0.913(0.004)	0.851(0.004)
		MCP	0.876(0.005)	0.961(0.002)	0.879(0.004)	0.933(0.003)
	1000	LASSO	0.814(0.005)	0.982(0.001)	0.720(0.009)	0.952(0.002)
		ENET	0.816(0.005)	0.978(0.001)	0.691(0.010)	0.951(0.002)
		SCAD	0.850(0.004)	0.981(0.000)	0.843(0.005)	0.968(0.001)
		MCP	0.815(0.004)	0.995(0.000)	0.805(0.004)	0.990(0.000)
1000	20	AIC	0.977(0.002)	0.817(0.012)	0.832(0.007)	0.842(0.012)
		BIC	0.916(0.003)	0.991(0.003)	0.982(0.002)	0.986(0.004)
		LASSO	0.988(0.002)	0.499(0.025)	0.999(0.001)	0.116(0.012)
		ENET	0.986(0.002)	0.479(0.025)	0.999(0.001)	0.083(0.010)
		SCAD	0.991(0.002)	0.595(0.022)	0.986(0.002)	0.503(0.022)
		MCP	0.980(0.003)	0.676(0.024)	0.976(0.003)	0.605(0.025)
	100	LASSO	0.966(0.004)	0.824(0.011)	0.985(0.002)	0.589(0.007)
		ENET	0.967(0.003)	0.810(0.012)	0.987(0.002)	0.509(0.007)
		SCAD	0.959(0.003)	0.901(0.004)	0.957(0.003)	0.875(0.005)
		MCP	0.941(0.004)	0.955(0.003)	0.930(0.004)	0.943(0.003)
	1000	LASSO	0.881(0.005)	0.983(0.001)	0.872(0.004)	0.945(0.001)
		ENET	0.863(0.005)	0.982(0.001)	0.864(0.005)	0.932(0.001)
		SCAD	0.917(0.004)	0.984(0.001)	0.909(0.004)	0.977(0.001)
		MCP	0.893(0.004)	0.995(0.000)	0.874(0.004)	0.994(0.000)
			AR(1) Covariance structure			
500	20	AIC	0.876(0.004)	0.797(0.015)	0.800(0.008)	0.785(0.014)
		BIC	0.789(0.003)	0.976(0.006)	0.972(0.004)	0.971(0.006)
		LASSO	0.952(0.004)	0.652(0.025)	0.979(0.003)	0.282(0.017)
		ENET	0.980(0.003)	0.617(0.027)	0.982(0.002)	0.231(0.016)
		SCAD	0.899(0.005)	0.688(0.020)	0.931(0.004)	0.549(0.022)
		MCP	0.863(0.006)	0.776(0.020)	0.919(0.005)	0.588(0.024)
	100	LASSO	0.936(0.004)	0.906(0.009)	0.970(0.003)	0.746(0.007)
		ENET	0.977(0.002)	0.907(0.008)	0.978(0.002)	0.704(0.007)
		SCAD	0.85(0.004)	0.902(0.003)	0.875(0.005)	0.863(0.004)
		MCP	0.794(0.004)	0.957(0.002)	0.826(0.005)	0.936(0.003)
	1000	LASSO	0.905(0.005)	0.989(0.001)	0.929(0.004)	0.964(0.001)
		ENET	0.950(0.004)	0.989(0.001)	0.950(0.004)	0.956(0.001)
		SCAD	0.755(0.005)	0.985(0.000)	0.803(0.004)	0.970(0.001)
		MCP	0.669(0.005)	0.995(0.000)	0.747(0.004)	0.991(0.000)
1000	20	AIC	0.934(0.004)	0.817(0.013)	0.816(0.008)	0.800(0.014)
		BIC	0.843(0.003)	0.982(0.005)	0.991(0.002)	0.988(0.004)
		LASSO	0.986(0.002)	0.639(0.025)	0.992(0.002)	0.302(0.018)
		ENET	0.996(0.001)	0.589(0.026)	0.994(0.001)	0.264(0.017)
		SCAD	0.941(0.004)	0.646(0.021)	0.953(0.004)	0.593(0.023)
		MCP	0.924(0.005)	0.724(0.021)	0.939(0.005)	0.638(0.025)
	100	LASSO	0.973(0.003)	0.885(0.009)	0.989(0.002)	0.745(0.006)
		ENET	0.995(0.001)	0.870(0.010)	0.992(0.002)	0.691(0.006)
		SCAD	0.900(0.004)	0.906(0.003)	0.908(0.004)	0.902(0.004)
		MCP	0.852(0.004)	0.959(0.002)	0.867(0.004)	0.949(0.003)
	1000	LASSO	0.959(0.003)	0.990(0.001)	0.970(0.003)	0.966(0.001)
		ENET	0.985(0.002)	0.992(0.001)	0.984(0.002)	0.956(0.001)
		SCAD	0.841(0.004)	0.980(0.001)	0.866(0.004)	0.978(0.001)
		MCP	0.786(0.004)	0.994(0.000)	0.815(0.004)	0.994(0.000)

The table reports the simulation mean (standard error) based on 200 iterations.

in which the tuning parameter $\lambda > 0$ controls a trade-off between the likelihood (or loss function) and the penalty. The tuning parameter can be determined by many different methods such as BIC

and cross-validation. The shrinkage parameter γ determines the degree of shrinkage of parameters in the concave penalty methods, which is either fixed or estimated. A similar setting of the penalized likelihood in (3.1) was introduced for zero-inflated Poisson model (Tang *et al.*, 2014). The structure of the two penalty functions, $P_1(\boldsymbol{\alpha}_{\mathcal{A}_1})$ and $P_2(\boldsymbol{\beta}_{\mathcal{A}_2})$, are the same except the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Hence, we discuss only $P_2(\boldsymbol{\beta}_{\mathcal{A}_2})$ in this article. A type of the penalty function is the bridge penalty function which has the following functional form

$$P_2\left(\boldsymbol{\beta}_{\mathcal{A}_2}\right) = \lambda \sum_{j\in\mathcal{A}_2} |\beta_j|^\gamma, \tag{3.2}$$

where the value of γ is related to famous regression methods. When $\gamma = 2$ in penalty (3.2), the penalized likelihood in (3.1) becomes ridge regression which was proposed to solve multicollinearity among predictors in the 1970s. Unfortunately, the ridge regression does not achieve variable selection because the parameter estimate could be close to zero but never achieves the zero value. When $\gamma = 1$, (3.1) is called LASSO regression, which simultaneously achieves variable selection and parameter estimation. ENET regression proposed by Zou and Hastie (2006) incorporated a convex combination of ridge and LASSO penalties to take advantage of both methods.

$$P_2\left(\boldsymbol{\beta}_{\mathcal{A}_2}\right) = \lambda_1 \sum_{j\in\mathcal{A}_2} |\beta_j| + \lambda_2 \sum_{j\in\mathcal{A}_2} \beta_j^2 = (1-\delta) \sum_{j\in\mathcal{A}_2} |\beta_j| + \delta \sum_{j\in\mathcal{A}_2} \beta_j^2,$$

where $\delta = \lambda_1/(\lambda_1 + \lambda_2)$.

Although the LASSO regression has substantial advantages, this method suffers from biased estimate under a certain condition (Zou, 2006). The solution of the bridge penalty is only continuous if $\gamma \geq 1$, and sparse only if $\gamma = 1$. When $\gamma = 1$, the LASSO solution is shifted by a constant value of λ. In order to simultaneously satisfy mathematical conditions for unbiasedness, sparsity, and continuity, non-convex penalty functions were introduced. We consider two important non-convex methods, SCAD (Fan and Li, 2001) and MCP (Zhang, 2010). The SCAD penalty function is given

$$P(\lambda, \gamma; |\beta_j|) = \begin{cases} \lambda|\beta_j|, & \text{if } |\beta_j| \leq \lambda, \\ \dfrac{2\gamma\lambda|\beta_j| - (\beta_j^2 + \lambda^2)}{2(\gamma-1)}, & \text{if } \lambda < |\beta_j| \leq \gamma\lambda, \\ \dfrac{\lambda^2(\gamma^2-1)}{2(\gamma-1)}, & \text{if } |\beta_j| > \gamma\lambda, \end{cases} \tag{3.3}$$

for $\lambda \geq 0$ and $\gamma > 2$. The MCP penalty is given:

$$P(\lambda, \gamma; |\beta_j|) = \begin{cases} \dfrac{2\lambda|\beta_j| - \beta_j^2}{2\gamma}, & \text{if } |\beta_j| \leq \gamma\lambda, \\ \dfrac{\gamma\lambda^2}{2}, & \text{if } |\beta_j| > \gamma\lambda, \end{cases} \tag{3.4}$$

for $\lambda \geq 0$ and $\gamma > 1$.

The original LASSO solution employed quadratic programming in Tibshirani (1996). Efron *et al.* (2004) demonstrated that the LASSO solution is a variant of the Least Angle Regression (LARS) algorithm, which is a computationally efficient stagewise procedure based on piecewise linearity. The coordinate descent algorithm proposed by Wu and Lange (2008) is significantly faster than LARS. This algorithm is incorporated in glmnet and ncvreg packages in R, where the former is for LASSO-type estimators, and the latter is for non-convex penalty estimators. Nonconvexity bares a burden of

numerical optimization, but the coordinate descent algorithm provides faster, stable numerical solutions.

The regularized regression method simultaneously achieves variable selection and parameter estimation, which are encompassed in the oracle property that an oracle estimator must hold asymptotic consistency in both variable selection and parameter estimation. Fan and Li (2001) and Zou (2006) posited that a good selection and estimation procedure should hold these two oracle properties. If $\lambda_n/n \to 0$ and $\lambda_n/\sqrt{n} \to \infty$, the LASSO estimator achieves a correct selection of the set of true non-zero parameters. However, Zou (2006) showed that when $\lambda_n = O(\sqrt{n})$, $\limsup_n P(\hat{\mathcal{A}} = \mathcal{A}) < 1$. To overcome this issue, Zou (2006) introduced the adaptive LASSO (ALASSO) under the assumption that $\lambda_n/\sqrt{n} \to 0$, which satisfies the oracle property. The SCAD and MCP penalty methods also achieve the oracle property.

Numerical optimization is a serious challenge in the regularized regression. Tibshirani (1996) used quadratic programming (QP) with a convex constraint as a special case of convex optimization to find the LASSO solution. The LARS algorithm is another technique to find the piecewise linear path, where the LARS algorithm is a homotopy method in the sense that the piecewise linear path is sequentially constructed. Meanwhile, SCAD and MCP encountered more serious challenge due to the nonconvexcity of penalties. The local linear approximation (LLA) algorithm using the LARS algorithm was proposed to find the solution for SCAD and MCP (Zou and Li, 2008). Regardless of the convexity or nonconvexity of the penalties, the CDA method was proposed as a fast and computationally efficient optimization method. The CDA method achieves the optimization of an objective function for a single parameter with fixing all other parameters and iteratively cycling through all parameters until convergence is achieved. The computational efficiency of CDA is $O(np)$, which is even better than the linear regression with QR decomposition ($O(np^2)$). Detailed rationale behind the CDA approach to MCP and SCAD is available in Breheny and Huang (2011). In this study, we incorporated the CDA method to compare the regularized regression methods.

The goal of parameter estimation is to achieve an (asymptotically) unbiased and consistent estimator. However, the goal of variable selection is more complex. The possible goals of variable selection include sensitivity, specificity, predictability, and selection consistency (Dziak *et al.* ,2012). Sensitivity measures how a method accurately includes the set of predictors with non-zero coefficients. Specificity measures how a method accurately excludes the set of predictors with zero coefficients. Predictability measures a certain degree of prediction error. Selection consistency indicates how a set of selected predictors approaches to the true set of predictors with non-zero coefficients as $n \to \infty$. As pointed out in many studies (Ng, 2013 and references therein), these four goals may not be achieved simultaneously in the sense that a high sensitivity (or specificity) does not guarantee a high predictability, and vice versa. Therefore, we compare sensitivity, specificity, and MSPE in the subsequent simulation study.

4. Simulation studies

Our simulation study aimed to investigate the performance of selected methods described in the previous section under various situations. The selected methods included LASSO, ENET, SCAD, and MCP. If $p = 100$, $2^p \approx 1.27 \times 10^{30}$ in which subset selection based on an information criterion is almost prohibited due to computational complexity. LASSO and ENET were implemented in the R glmnet package (Friedman *et al.*, 2009) and SCAD and MCP were implemented in the R ncvreg package (Breheny, 2013).

The design of the simulation study for the two-part model considered the following issues. First,

Table 4: Prediction performance of parameter space P2 for two covariance structures

n	p	Method	Independent covariance structure				
			Mean squared error		Classification		
			Positive	All	Accuracy	Sensitivity	Specificity
500	20	AIC	30.26(0.31)	31.04(0.31)	0.890(0.001)	0.878(0.002)	0.899(0.001)
		BIC	29.89(0.32)	30.47(0.32)	0.888(0.001)	0.875(0.002)	0.897(0.001)
		LASSO	32.79(0.26)	34.10(0.25)	0.888(0.001)	0.880(0.002)	0.894(0.001)
		ENET	32.79(0.26)	34.14(0.25)	0.889(0.001)	0.885(0.002)	0.892(0.001)
		SCAD	33.41(0.26)	34.74(0.26)	0.889(0.001)	0.876(0.002)	0.899(0.001)
		MCP	35.00(0.25)	34.77(0.25)	0.889(0.001)	0.876(0.002)	0.899(0.001)
	100	LASSO	30.81(0.30)	31.90(0.29)	0.878(0.001)	0.874(0.002)	0.881(0.002)
		ENET	30.74(0.30)	31.82(0.29)	0.875(0.001)	0.879(0.002)	0.873(0.002)
		SCAD	33.80(0.30)	35.29(0.30)	0.883(0.001)	0.868(0.002)	0.895(0.001)
		MCP	35.51(0.30)	35.20(0.31)	0.884(0.001)	0.870(0.002)	0.895(0.001)
	1000	LASSO	30.74(0.26)	31.44(0.26)	0.844(0.001)	0.864(0.002)	0.834(0.002)
		ENET	30.58(0.27)	31.22(0.26)	0.836(0.001)	0.878(0.003)	0.815(0.002)
		SCAD	33.66(0.23)	35.14(0.23)	0.862(0.001)	0.847(0.002)	0.873(0.002)
		MCP	34.81(0.21)	34.81(0.22)	0.864(0.001)	0.850(0.002)	0.875(0.002)
1000	20	AIC	29.62(0.24)	30.22(0.24)	0.897(0.001)	0.882(0.001)	0.908(0.001)
		BIC	29.59(0.24)	30.15(0.24)	0.896(0.001)	0.881(0.001)	0.908(0.001)
		LASSO	31.62(0.21)	32.97(0.20)	0.897(0.001)	0.886(0.001)	0.904(0.001)
		ENET	31.62(0.21)	32.97(0.20)	0.896(0.001)	0.889(0.001)	0.902(0.001)
		SCAD	31.95(0.20)	33.33(0.20)	0.897(0.001)	0.882(0.001)	0.908(0.001)
		MCP	33.50(0.19)	33.30(0.20)	0.897(0.001)	0.882(0.001)	0.908(0.001)
	100	LASSO	30.05(0.23)	30.95(0.22)	0.886(0.001)	0.875(0.001)	0.894(0.001)
		ENET	30.09(0.23)	30.99(0.22)	0.885(0.001)	0.880(0.001)	0.889(0.001)
		SCAD	30.84(0.21)	32.04(0.19)	0.890(0.001)	0.875(0.001)	0.902(0.001)
		MCP	31.97(0.18)	32.07(0.19)	0.890(0.001)	0.875(0.001)	0.902(0.001)
	1000	LASSO	28.77(0.20)	29.40(0.19)	0.872(0.001)	0.873(0.001)	0.872(0.001)
		ENET	28.78(0.20)	29.38(0.20)	0.868(0.001)	0.883(0.001)	0.859(0.001)
		SCAD	30.08(0.15)	31.35(0.14)	0.881(0.001)	0.864(0.001)	0.894(0.001)
		MCP	31.34(0.13)	31.24(0.14)	0.882(0.001)	0.865(0.001)	0.895(0.001)
			AR(1) covariance structure				
500	20	AIC	45.94(0.43)	47.15(0.44)	0.914(0.001)	0.907(0.001)	0.920(0.001)
		BIC	43.84(0.35)	44.65(0.36)	0.912(0.001)	0.906(0.001)	0.918(0.001)
		LASSO	46.37(0.31)	47.97(0.31)	0.914(0.001)	0.907(0.001)	0.921(0.001)
		ENET	46.22(0.31)	47.92(0.31)	0.915(0.001)	0.910(0.001)	0.921(0.001)
		SCAD	47.20(0.31)	48.92(0.31)	0.913(0.001)	0.906(0.001)	0.919(0.001)
		MCP	49.08(0.30)	48.93(0.31)	0.913(0.001)	0.907(0.001)	0.919(0.001)
	100	LASSO	46.56(0.44)	47.86(0.42)	0.911(0.001)	0.910(0.002)	0.912(0.002)
		ENET	46.42(0.43)	47.76(0.42)	0.911(0.001)	0.913(0.002)	0.911(0.002)
		SCAD	49.63(0.49)	51.37(0.49)	0.906(0.001)	0.900(0.002)	0.911(0.002)
		MCP	51.56(0.50)	51.36(0.50)	0.907(0.001)	0.900(0.002)	0.912(0.002)
	1000	LASSO	42.57(0.27)	43.49(0.26)	0.904(0.001)	0.912(0.002)	0.899(0.002)
		ENET	42.40(0.28)	43.27(0.26)	0.906(0.001)	0.920(0.002)	0.896(0.002)
		SCAD	47.06(0.25)	48.68(0.25)	0.890(0.001)	0.886(0.002)	0.894(0.002)
		MCP	48.28(0.24)	48.19(0.24)	0.892(0.001)	0.888(0.002)	0.896(0.002)
1000	20	AIC	45.49(0.39)	46.30(0.39)	0.922(0.001)	0.913(0.001)	0.930(0.001)
		BIC	45.16(0.39)	45.76(0.39)	0.921(0.001)	0.911(0.001)	0.930(0.001)
		LASSO	49.20(0.41)	50.71(0.41)	0.922(0.001)	0.916(0.001)	0.927(0.001)
		ENET	49.18(0.40)	50.7(0.41)	0.922(0.001)	0.918(0.001)	0.925(0.001)
		SCAD	49.79(0.41)	51.36(0.41)	0.922(0.001)	0.912(0.001)	0.930(0.001)
		MCP	51.33(0.4)	51.37(0.41)	0.922(0.001)	0.912(0.001)	0.930(0.001)
	100	LASSO	44.13(0.24)	45.43(0.22)	0.927(0.001)	0.924(0.001)	0.930(0.001)
		ENET	44.07(0.24)	45.32(0.23)	0.927(0.001)	0.927(0.001)	0.927(0.001)
		SCAD	46.24(0.21)	47.82(0.20)	0.926(0.001)	0.917(0.001)	0.933(0.001)
		MCP	47.91(0.19)	47.84(0.19)	0.926(0.001)	0.917(0.001)	0.933(0.001)
	1000	LASSO	45.54(0.29)	46.65(0.28)	0.914(0.001)	0.909(0.001)	0.919(0.001)
		ENET	45.42(0.30)	46.49(0.28)	0.914(0.001)	0.912(0.001)	0.916(0.001)
		SCAD	48.74(0.24)	50.41(0.23)	0.910(0.001)	0.900(0.001)	0.918(0.001)
		MCP	50.57(0.22)	50.33(0.23)	0.912(0.001)	0.902(0.001)	0.920(0.001)

The table reports the simulation mean (standard error) based on 200 iterations.

we considered two different sample sizes, $n = 500$, and 1000, for both the training and test data, which means that the ratio of both the data is one to one. The different sample sizes allowed to check the

behavior of variable selection and prediction performance as sample size varied. Second, the length of the parameter space, or the number of predictors, are $p = 20, 100, 1,000$ without the intercept. We consider two parameter spaces as follows:

$$P1 : (3, 1.5, 0, 0, 2, \underbrace{0, \ldots, 0}_{p-5})$$

$$P2 : (\underbrace{1.51, 1.78, 1.58, 1.80, 1.05, 0.54, 0.82, 1.14, 1.52, 1.98, 0.31, 0.52, 1.54, 0.25, 0.93}_{15}, \underbrace{0, \ldots, 0}_{p-15}).$$

The first parameter space comprises few strong coefficients, which was used in the seminal LASSO paper (Tibshirani, 1996) and many other studies (Fan and Li, 2005; Zou and Hastie, 2001). The second parameter space was introduced to investigate how the selected penalized regression methods perform for diverse coefficient values.

Third, the covariance structure among predictors is one of the most important factors affecting the variable selection performance. We considered the independent and AR(1) covariance structures as follows:

1. Independence: $\Sigma_x = \sigma^2 I_p$.
2. AR(1) correlation: $\sigma^2 \sigma_{ij} = \rho^{|i-j|}$ for $i = 1, \ldots, p$ and $j = 1, \ldots, p$.

Our simulation study used that $\sigma = 3$ and $\rho = 0.5$. Fourth, all predictors were standardized after being generated from a multivariate normal distribution, which was implemented in the R mvrnorm package. Fifth, in TPM, LM and GLM share the same data set where LM uses a less portion of the data. Hence, the proportion of zero values could affect the performance. In our case, we set 50–60 % data positive by controlling the intercept value in the simulation because a small portion of zeros or positive values can cause biased estimation in logistic regression. The shrinkage parameter (γ) in SCAD and MCP used the default value in the package, and the tuning parameter was estimated using a 10-fold cross-validation in the training data. All other options used the default values in the packages.

We focus on the two types of performance evaluation: Variable selection and prediction performances. These performance measures are closely related to the oracle properties. The variable selection performance is measured by β-sensitivity and β-specificity, which are defined as:

$$\beta - \text{Sensitivity} = \frac{1}{R} \sum_{r=1}^{R} I\left(\hat{\mathcal{A}} = \mathcal{A}\right), \tag{4.1}$$

$$\beta - \text{Specificity} = \frac{1}{R} \sum_{r=1}^{R} I\left(\hat{\mathcal{A}}^c = \mathcal{A}^c\right), \tag{4.2}$$

where $\mathcal{A}$ is either $\mathcal{A}_1$ or $\mathcal{A}_2$ described in section 2, $\hat{\mathcal{A}} = \{j : \hat{\theta}_j \neq 0\}$ and $\hat{\mathcal{A}}^c = \{j : \hat{\theta}_j = 0\}$ for the estimated parameter space $\hat{\theta}$ using any selected method, and R indicates the number of iterations, which is 200 in our study. The β-sensitivity measures how accurately a selection method includes the predictors with non-zero coefficients in the model, and the β-specificity measures how accurately a selection method exclude the predictors with zero coefficients from the model.

The prediction performance for the GLM and LM is measured separately using accuracy, sensitivity, and specificity as metrics for the GLM, and the mean squared prediction error (MSPE) for the LM.

The GLM metrics are used for the classification between zero and positive values. The mean absolute deviation serves as an alternative to the MSPE. These metrics are all evaluated in the testing data set. It is considered that the lower the MSPE, the better the prediction performance of the LM, while for the GLM, values of accuracy, sensitivity, and specificity closer to one indicates better classification performance.

Tables 1–4 reported the average and standard error of the performance measures obtained from 200 simulations to evaluate variable selection and prediction performance in the test data set. Table 1 presented the average values of β-sensitivity and β-specificity for the logistic regression (GLM) and linear regression (LM) for the parameter space P1. Table 2 presented the two MSPE values for the positive values and all values of y and the classification measures such as accuracy, sensitivity, and sprecificity for the binary responses in the GLM. Similarly, we presented the variable selection and prediction performance for the paramter space P2 in Tables 3 and 4, respectively. The performance results for AIC and BIC were only reported for the small number of predictors, $p = 20$.

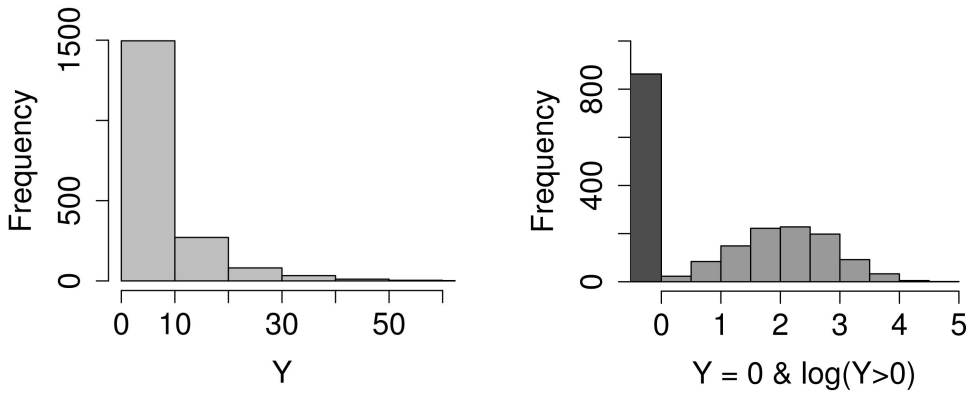
As evident in Table 1, all the methods achieved better β-sensitivity average values although they showed various results in β-specificity for the parameter space P1, which contains strong coefficient values. The AR(1) covariance strucucture showed slightly lower performance than the independent covariance structure. In Table 2, the prediction results showed slightly different performance among the selected methods. For the average test MSPE, the LASSO and ENET methods outperformed the SCAD and MCP methods. However, the methods demonstrated no clear differences in classification between the zero and positive values. All the standard error values had no significant differences across the selected methods.

For the parameter space P2, which contains coefficients generated from the uniform distribution between 0.5 and 2, the selected methods overall presented underperformance in variable selection and prediction assessment, as can be seen Tables 3 and 4. Unlike the results of P1, P2 demonstrated that the LASSO and ENET methods outperform the SCAD and MCP methods in terms of β-sensitivity.

Our simulation results showed that penalized TPMs achieve better performance in both variable selection and prediction in high-dimensional data, regardless of covariance structures and sample sizes. The selected nonconvex methods such as SCAD and MCP exhibited better performance in β-specificity than the LASSO-type methods. The LASSO-type methods exhibited better performance in prediction than the nonconvex methods, regardless of the parameter spaces. Meanwhile, traditional variable selection methods based on AIC and BIC achieved better performance in β-specificity and MSPE than penalized methods for low-dimensional data.

5. Empirical studies

Our empirical study for TPM was conducted using community-based crime data. Crime data can be collected in many ways. One common way is to collect crime data for each community such as city or town. Each community possesses its idiosyncratic characteristics with respect to demographics and socioeconomic status. Redmond and Baveja (2002) generated a comprehensive community-based crime data which is available at UC Irvine Machine Learning Repository(https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime). The data set contains 2,215 observations (communities), 124 predictors, and 18 response variables (9 different types of crimes with the original frequency and the frequency per 100,000 inhabitants) from multiple original data sources such as socio-economic data from the 1990 Census, law enforcement data from the 1990 Law Enforcement Management and Administrative Statistics (LEMAS), and crime data from the 1995 FBI Uniform Crime Report (UCR) Statistics.



(a) *Original semicontinuous data* (b) *Zero values and log-transformed data*

Figure 1: *Semicontinuous distribution of murder per 100,000 habitants.*

The Census data include age-, race-, income-, family-, and house-related variables. The UCR data contained the original counts and the counts per 100,000 inhabitants for murder, rape, robbery, assault, burglary, larceny, auto theft, and arson. The LEMAS data collected policing-related data from state and local law enforcement agencies, including all those that employ 100 or more sworn police officers and a nationally representative sample of smaller agencies. Many communities with a small number of sworn officers had missing values for the predictors from the LEMAS data.

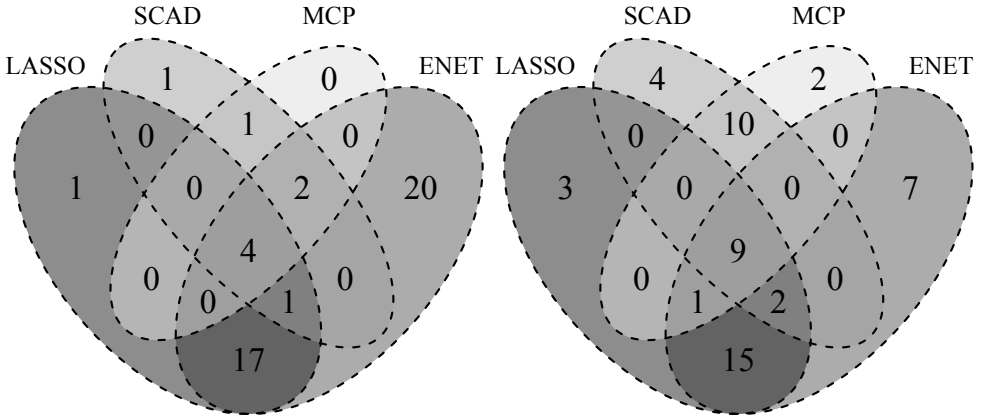
Among the 9 types of crime, murder, arson, and rape showed semicontinuous distributions, where these types of crime contained a considerable number of zero values over 10% of the total number of observations and demonstrated a skewed right distribution as the mean is much greater than the median. This empirical study focused on the response variable of the murder incidents per 100,000 habitants of which semicontinuous property is demonstrated in Figure 1. In Figure 1, the right-side figure illustrated the zero values and a bell-shaped curve of log-transformed positive values, which was modeled via the TPM. After removing the 23 LEMAS variables and any communities with significant amount of missing values, our final analytical data set consisted of one dependent variable and 101 predictors in 1,901 communities.

In order to check multicollinearity among 101 predictors, first we identified the perfect linearity (and hence, multicollinearity) between OwnOccQrange and OwnOccLowQuart/OwnOccHiQuart as well as RentQrange and RentLowQ/RentHighQ using the alias function for the lm function in R. Therefore, we removed two variables, OwnOccQrange and RentQrange. Using the variance influence factor (VIF), we examined multicollinearity among 99 predictors. The variance inflation factor (VIF) analysis shows that the 83 out of 99 predictors had the squared VIF value greater than 2, which indicates that the community-based crime data presented severe multicollinearity among predictors. In summary, the community-based crime data was characterized by skewed-right responses with semicontinuity and a fairly large number of predictor with multicollinearity. We identified a set of predictors via various variable selection methods and evaluated their prediction performance described in the previous sections. We split the whole data into the training and test data sets with $1:1$ ratio, which resulted in a slightly different sample sizes for the linear regression in the training and test data

Table 5: MSPE and the number of selected predictors

	Sample size		Selection Methods					
	Train	Test	AIC	BIC	LASSO	ENET	SCAD	MCP
GLM	941	941	30	10	23	44	9	7
LM	492	530	38	4	30	34	25	22
MSPE			77.48	61.15	45.28	44.74	43.26	43.37

* LASSO, least absolute shrinkage and selection operator; ENET, Elastic NeT; MCP, Minimax Concave Penalty; SCAD, Smoothly Clipped absolute Deviation; MSPE, Mean squared prediction error; The integer values in Selection Methods denote the number of predictors selected by each method.



(a) *Penalized GLM* (b) *Penalized LM*

Figure 2: *Venn diagram of predictors selected by four penalized methods.*

sets.

In Table 5, we demonstrated the MSPE and the number of predictors selected by each methods. This result should be cautiously interpreted because a different sampling of training and test data may lead to a different result. Overall, the folded concave penalty methods of SCAD and MCP outperformed other methods with respect to the MSPE. The MSPE of MCP is 12% lower than the ENET one. AIC and ENET methods tend to select a higher number of predictors, and BIC and MCP tend to select a smaller number of predictors, which is consistent to the simulation results reported in the previous section.

In Figure 2, we presented two Venn diagrams of predictors selected by the four penalized methods from both GLM and LM in Table 5 to closely look at the patterns among the selected predictors. The LASSO predictors were mostly selected by the ENET method, and the MCP predictors were mostly selected by SCAD. The LASSO and ENET selected more predictors than SCAD and MCP in both models. The four penalized methods in both GLM and LM commonly selected seven predictors, although their compositions were different. The commonly selected variables from the logistic regression among four methods included population, blackPerCap, PersPerOwnOccHous, PersPerRentOccHous, MedNumBR, LandArea, and racePctWhite, and the commonly selected variables from the linear regression among four methods included PersPerFam, NumKidsBornNeverMar, NumImmig,

PersPerOwnOccHous, PopDens, racePctWhite, and PctWorkMomYoungKids where the two predictors, PersPerOwnOccHous(mean persons per owner occupied household) and racePctWhite (percentage of population that is Caucasian) were common for both models. For further information, refer to the dictionary of these predictors at the UCI Machine Learning Repository mentioned above.

6. Discussion

In this study, we investigated penalized regression-based variable selection methods for two-part models. We conducted simulation studies under diverse statistical assumptions and an empirical study using community-based crime data. Our analytical results demonstrated that penalized TPMs achieve better performance in both variable selection and prediction in high-dimensional data, regardless of covariance structures and sample sizes. Moreover, the LASSO-type methods such as LASSO and ENET outperformed the nonconvex methods such as SCAD and MCP in mean squared error. In simulation studies, for a small number of predictors, for example, $p = 20$, traditional variable selection methods based on information criteria achieved better performance in β-specificity and MSE than penalized methods.

Variable selection in the TPM is affected by several unique features in addition to conventional matters, such as high dimensionality, multicollinearity, covariance structure, and sparsity. These TPM-specific features include the same pool of predictors for both models in (2.2) and (2.3), a smaller sample size for modeling positive values, and the mean squared error as a function of the probability of a positive value and the marginal mean of the positive values. Our simulation was designed such that positive values were 50–60% of the total sample size. Simulation results showed that the sensitivity of LM was as good as those of GLM while the specificity of LM was similar to or worse than those of GLM. As can be seen in (2.8), the probability of a positive value results in a smaller MSE for TPM compared to that of the linear regression for the positive values.

Additionally, the simulation study also showed that the best variable selection may not be associated with the least prediction error. The trade-off between consistent variable selection and efficient prediction is well addressed in Ng (2013) and references therein. The convex penalty methods and the folded concave penalty methods show different behaviors in sensitivity and specificity. When the coefficients are quite different from zero as in Table 1, both penalty methods performed well for sensitivity. On the other hand, when coefficients were generated from Uniform[0.25, 2] as in Tables 3 and 4, the two methods exhibited similar sensitivities, or the convex methods marginally performed better. The folded concave methods outperformed the convex penalty methods in the specificity performance regardless of parameter spaces and covariance structures. This result is partly because the convex penalty methods are inclined to include more predictors with non-zero coefficient estimates, as explained in Breheny and Huang (2011).

Our current study can be extended in several directions. First, as we only considered the selected number of methods, it is worthwhile to consider recently developed variable selection methods especially for high dimensional data such as interaction selection (Hao *et al.*, 2018) and screening methods (Fan and Lv, 2008; Pan *et al.*, 2019; Tibshirani *et al.*, 2012). In particular, screening out insignificant predictors is expected to help improve the specificity issue of penalized regression methods. Second, the positive values of a semicontinuous variable might follow the log-skewed normal distribution, necessiating a generalized regression beyond a linear regression under the normality assumption. It is desirable to consider regularized GLM with gamma or skewed lognormal distribution for various degrees of positive skewness.

References

Breheny P (2013). ncvreg: Regularization paths for scad-and mcp-penalized regression models, R package version, 2.6-0, Available from: https://pbreheny.github.io/ncvreg/

Breheny P and Huang J (2011). Coordinate descent algorithms for nonconvex penalized regression with applications to biological feature selection, *The Annals of Applied Statistics*, **5**, 232–253.

Brown EC, Catalano RF, Fleming CB, Haggerty KP, and Abbott RD (2005). Adolescent substance use outcomes in the raising healthy children project: A two-part latent growth curve analysis, *Journal of Consulting and Clinical Psychology*, **73**, 699–710.

Candes E and Tao T (2007). The Dantzig selector: Statistical estimation when p is much larger than n, *The Annals of Statistics*, **35**, 2313–2351.

Cragg JG (1971). Some statistical models for limited dependent variables with application to the demand for durable goods, *Econometrica: Journal of the Econometric Society*, **39**, 829–844.

Duan N, Manning WG, Morris CN, and Newhouse JPA (1983). Comparison of alternative models for the demand for medical care, *Journal of Business and Economic Statistics*, **1**, 115–126.

Dunn PK and Smyth GK (2005). Series evaluation of Tweedie exponential dispersion model densities, *Statistics and Computing*, **15**, 267–280.

Dziak JJ, Coffman DL, Lanza ST, and Li R (2020). Sensitivity and specificity of information criteria, *Briefings in Bioinformatics*, **21**, 553–565.

Efron B, Hastie T, Johnstone I, and Tibshirani R (2004). Least angle regression, *The Annals of Statistics*, **32**, 407–451.

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.

Fan J and Lv J (2008). Sure independence screening for ultrahigh dimensional feature space, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**, 849–911.

Frees EW, Jin X, and Lin X (2013). Actuarial applications of multivariate two-part regression models, *Annals of Actuarial Science* **7**, 258–287.

Friedman J, Hastie T, and Tibshirani R (2009). glmnet: Lasso and elastic-net regularized generalized linear models, R package version, 1.0, Available from: https://cran.r-project.org/web/packages/glmnet

Hao N, Feng Y, and Zhang HH (2018). Model selection for high-dimensional quadratic regression via regularization, *Journal of the American Statistical Association*, **113**, 615–625.

Kang HW and Kang HB (2017). Prediction of crime occurrence from multi-modal data using deep learning, *PloS One* **12**, e0176244.

Kokonendji CC, Bonat WH, and Abid R (2021). Tweedie regression models and its geometric sums for (semi-) continuous data, *Wiley Interdisciplinary Reviews: Computational Statistics*, **13**, e1496.

Liu L (2009). Joint modeling longitudinal semi-continuous data and survival with application to longitudinal medical cost data, *Statistics in Medicine*, **28**, 972–986.

Merlo L, Maruotti A, and Petrella L (2022). Two-part quantile regression models for semi-continuous longitudinal data: A finite mixture approach, *Statistical Modelling*, **22**, 485–508.

Min Y and Agresti A (2002). Modeling nonnegative data with clumping at zero: A survey, *Journal of the Iranian Statistical Society*, **1**, 7–33.

Mullahy J (1998). Much ado about two: Reconsidering retransformation and the two-part model in health econometrics, *Journal of Health Economics*, **17**, 247–281. Notice: Data not available: U.S. Bureau of Labor Statistics (n.d.).

Neelon B, O'Malley AJ, and Smith VA (2016). Modeling zero-modified count and semicontinuous data in health services research Part 1: Background and overview, *Statistics in Medicine*, **35**, 5070–5093.

Ng S (2013). Variable selection in predictive regressions, In *Handbook of Economic Forecasting*; Elliott G and Timmermann A, Eds, Elsvier, 752–789.

Olsen MK and Schafer JL (2001). A two-part random-effects model for semicontinuous longitudinal data, *Journal of the American Statistical Association*, **96**, 730–745.

Pan W, Wang X, Xiao W, and Zhu H (2019). A generic sure independence screening procedure, *Journal of the American Statistical Association*, **114**, 928–937.

Redmond MA and Baveja A (2002). A data-driven software tool for enabling cooperative information sharing among police departments, *European Journal of Operational Research*, **141**, 660–678.

Smith VA, Preisser JS, Neelon B, and Maciejewski ML (2014). A marginalized two-part model for semicontinuous data, *Statistics in Medicine*, **33**, 4891–4903.

Tang Y, Xiang L, and Zhu Z (2014). Risk factor selection in rate making: EM adaptive LASSO for zero-inflated poisson regression models, *Risk Analysis*, **34**, 1112–1127.

Tibshirani R (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**, 267–288.

Tibshirani R, Bien J, Friedman J, Hastie T, Simon N, Taylor J, and Tibshirani RJ (2012). Strong rules for discarding predictors in lasso-type problems, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **74**, 245–266.

Tu W and Zhou XHA (1999). Wald test comparing medical costs based on log-normal distributions with zero valued costs, *Statistics in Medicine*, **18**, 2749–2761.

Tweedie MCK (1984). An index which distinguishes between some important exponential families, Statistics: Applications and New Directions, In Ghosh JK and Roy J (Eds), *Indian Statistical Institute, Calcutta*, 579–604.

Wu TT and Lange K (2008). Coordinate descent algorithms for lasso penalized regression, *The Annals of Applied Statistics*, **2**, 224–244.

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

Zhang C-H (2010). Nearly unbiased variable selection under minimax concave penalty, *The Annals of statistics*, **38**, 894–942.

Zhao T, Luo X, Chu H, Le CT, Epstein LH, and Thomas JL (2016). A two-part mixed effects model for cigarette purchase task data, *Journal of the Experimental Analysis of Behavior*, **106**, 242–253.

Zou B, Mi X, Xenakis J, Wu D, Hu J, and Zou F (2023). A deep neural network two-part model and feature importance test for semi-continuous data, *bioRxiv*, 2023-06, Available from: https://doi.org/10.11 01/2023.06.07.544106

Zou H (2006). The adaptive lasso and its oracle properties, *Journal of the American Statistical Association*, **101**, 1418–1429.

Zou H and Hastie T (2005). Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301–320.

Zou H and Li R (2008). One-step sparse estimates in nonconcave penalized likelihood models, *The Annals of Statistics*, **36**, 1509–1533.

Received January 07, 2024; Revised May 14, 2024; Accepted May 14, 2024