Thailand Statistician
July 2020; 18(3): 306-318
http://statassoc.or.th
Contributed paper

Performance Comparison of Penalized Regression Methods in Poisson Regression under High-Dimensional Sparse Data with Multicollinearity

Chutikarn Choosawat, Orawan Reangsephet, Patchanok Srisuradetchai and Supranee Lisawadi*
Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Rangsit Campus, Pathumthani, Thailand.
*Corresponding author; e-mail: supranee@mathstat.sci.tu.ac.th

Received: 27 September 2018
Revised: 2 January 2019
Accepted: 3 July 2019

Abstract

Ridge regression, least absolute shrinkage and selection operator (LASSO), and adaptive LASSO can be employed for fitting high-dimensional count data by using the Poisson model. However, the performance of these statistical models has not been explicitly studied under the condition of sparse data with a multicollinearity problem. Thus, this paper aims to study the performance and compare ridge regression, LASSO and adaptive LASSO by using the criteria of median prediction mean square error (mPMSE), False Negative Rate (FNR), and False Positive Rate (FPR). The correlation structures of constant, Toeplitz, and hub Toeplitz are considered. Monte Carlo simulations with 1,000 iterations were performed to achieve the goal. The results showed that adaptive LASSO produced the lowest mPMSE. When the correlation was higher, ridge regression had the lowest mPMSE. Two criteria of incorrect variable selection were analyzed (FNR and FPR). In terms of FNR, LASSO performed better than adaptive LASSO. In terms of FPR, the opposite was true. We carried out simulations to examine the performance of the mPMSE for ridge regression, LASSO, and adaptive LASSO. We also compared the variable selection of LASSO and adaptive LASSO, using two criteria of incorrect variable selection (FNR and FPR).

Keywords: Ridge regression, LASSO, adaptive LASSO, Monte Carlo simulation.

1. Introduction

The standard statistical method for analyzing count data is the Poisson regression model, which studies the relationship between the mean of count data and explanatory variables. In reality, many explanatory variables can be correlated in a certain degree. This creates a problem called "multicollinearity". Also, nowadays a situation where the number of independent variables is larger than the number of observations can occur and it is called the "high dimensional data" problem. Very large datasets with increasing dimensions are being generated in many fields such as genetics, medicine, economics, engineering and social science. High-dimensional data have posed new

challenges to statistical analysis, because modeling introduces model overfitting, estimation instability, and computational difficulty. Many previous studies have used high and low dimensional linear models to apply penalized estimation to data analysis and to compare the performance of each estimator, for example, Pungpapong (2014), Oyeyemi et al. (2015), Ahmed and Yüzbaşı (2016), Yüzbaşı et al. (2017b), Gao et al. (2017), Yüzbası et al. (2017a).

In this paper, high-dimensional data in which the number of independent variables is greater than the sample size are of interest. Furthermore, independent variables can be highly correlated, but only a few of them effect the mean of count data. In such cases, ordinary least squares (OLS) and maximum likelihood estimation (MLE) might not provide a best solution. If the independent variables are highly correlated, the variance of the MLE is increased and interpretation of results can be difficult and complex. Hence, the MLE is not recommended when the independent variables are highly correlated and/or high dimensional.

Penalized regression is a popular methodology for high dimensional data to estimate regression coefficients. The estimated coefficients are derived by minimizing the objective function as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \left\| \mathbf{y} - \exp\left(\mathbf{X}\boldsymbol{\beta}\right) \right\|^2 + P_{\lambda}\left(\boldsymbol{\beta}\right),$$

where $P_{\lambda}(\boldsymbol{\beta})$ is a penalty function. There are many forms of penalty functions and the choice depends on the method of penalizing regression. This study investigates penalized regression using three methods: Ridge regression, least absolute shrinkage and selection operator (LASSO), and adaptive LASSO. These three methods can solve the multicollinearity problem as they can shrink the coefficients of regression. The statistical qualification of LASSO in the Poisson regression model was developed by Hossain and Ahmed (2012). This allows simultaneous coefficient estimation and variable selection by assigning the value zero to some independent variables. For this reason, LASSO is widely used for analysis of high-dimensional data; however, LASSO has some shortcomings. When there is multicollinearity, LASSO may select one variable among correlated variables, violating "consistency". Ivanoff et al. (2016) improved the efficiency of parameter estimation and variable selection in Poisson regression by proposing Adaptive LASSO, in which weights are used to penalize different coefficients. The ℓ_1-norm penalty improves variable selection and solves the problem of LASSO when multicollinearity occurs. Zou (2006) developed and presented a qualification of ridge regression in Poisson regression by estimating the regression coefficient using the ℓ_2-norm penalty. It is well-known that ridge regression shrinks the coefficient of independent variables but it cannot select variables for the model. Algamal and Lee (2015) proposed the adjusted adaptive LASSO estimator in high-dimensional Poisson regression with multicollinearity. They proposed an adjustment of adaptive LASSO to take into account the maximum likelihood standard errors of the coefficient parameters. Ivanoff et al. (2016) estimated the intensity function of the Poisson regression model by using a generalization of the classical basic approach, combined with the LASSO or the group-LASSO procedure. Selection depends on penalty weights that must be calibrated. They showed that the associated LASSO and group-LASSO procedures satisfy fast and slow oracle inequalities.

The "sparse data" means that there are too many non-significant predictors, in which their coefficients are zero, in the model. This is different from missing data, in which some or many of the values are unknown. In this paper, we will assess the performance of the three estimators in cases with different sample sizes and numbers of independent variables.

2. Methodologies

2.1. Poisson regression model

Suppose that Y is a Poisson distribution with a conditional mean (μ) which depends on the individual characteristics of Y

$$P(Y=y)=f(y)=\frac{\mu^{y}\exp(-\mu)}{y!}, \tag{1}$$

where $y=0,1,2,\ldots$. The conditional mean parameter in a Poisson distribution, $E(Y)=\mu$ and the variance parameter $\sigma^2(Y)=\mu$. However, each Y_i can have its own μ_i. Let $\mathbf{y}=(y_1,\ldots,y_n)^T$ be a vector of the response variables, $\boldsymbol{\beta}=(\beta_1,\beta_2,\ldots,\beta_p)^T$ be a vector of unknown coefficients, and $\boldsymbol{\mu}=(\mu_1,\ldots,\mu_n)^T$ is the vector of mean of the Poisson regression, the Poisson regression model with p available predictors for y_i is given by

$$f(y_i;\mu_i,\mathbf{X}_i)=\frac{\mu_i^{y_i}\exp(-\mu_i)}{y_i!}, \tag{2}$$

where $E(Y_i\mid\mathbf{X}_i)=V(Y_i\mid\mathbf{X}_i)=\mu_i=\exp(\mathbf{X}_i^T\boldsymbol{\beta})$ for $i=1,2,3,\ldots,n$.

Consider the following high-dimensional sparse data model

$$\log(E(Y_i\mid\mathbf{X}_i))=\mathbf{X}_i^T\boldsymbol{\beta},$$

where $i=1,2,\ldots,n$ and $p>n$. Under the sparsity assumption on $\boldsymbol{\beta}$, we assume that there are q non-significant predictors, where $q<p$. Hence, $\mathbf{X}_i$ can be decomposed as $\mathbf{X}_i=(\mathbf{x}_{iA},\mathbf{x}_{iB})$ with $\mathbf{x}_{iA}=(\mathbf{x}_{i1},\ldots,\mathbf{x}_{i(p-q)})^T\in\mathbb{R}^{p-q}$ and $\mathbf{x}_{iB}=(\mathbf{x}_{i(p-q+1)},\ldots,\mathbf{x}_{ip})^T\in\mathbb{R}^{q}$ where $p-q$ is smaller than the sample size n. Finally, let $\mathbf{X}=\left(\mathbf{X}_A,\mathbf{X}_B\right)^T\in\mathbb{R}^{n\times p}$ be the matrix that includes all predictive variables. Here, $\mathrm{X}_A=(\mathbf{x}_{iA},\ldots,\mathbf{x}_{nA})^T\in\mathbb{R}^{n\times(p-q)}$ and $\mathbf{X}_B=(\mathbf{x}_{iB},\ldots,\mathbf{x}_{nB})^T\in\mathbb{R}^{n\times q}$ are the matrices associated with $\mathbf{x}_{iA}$ and $\mathbf{x}_{iB}$, respectively.

Regression parameters can be estimated by MLE, which is based on probability. First, one must specify the likelihood function of the random variables and find the maximum value of this function. Under the assumption of independent observations, the log likelihood function is given by

$$l(\boldsymbol{\mu};\mathbf{y}) \;=\; \ln L(\boldsymbol{\beta}) \;=\; \sum_{i=1}^{n}\left(y_i\mathbf{X}_i^T\boldsymbol{\beta}-\exp(\mathbf{X}_i^T\boldsymbol{\beta})-\ln y_i!\right). \tag{3}$$

The derivative of the log-likelihood with respect to $\boldsymbol{\beta}$ is obtained by the chain rule

$$\frac{\partial}{\partial\boldsymbol{\beta}}\ln L(\boldsymbol{\beta}) \;=\; \sum_{i=1}^{n}\left(y_i-\exp(\mathbf{X}_i^T\boldsymbol{\beta})\right)\mathbf{X}_i=0. \tag{4}$$

The result of this equation is nonlinear form in $\boldsymbol{\beta}$, and $\hat{\boldsymbol{\beta}}$ can be solved using an iterative method such as the Newton-Raphson. Månsson and Shukur (2011) introduced the following iterative weighted least square (IWLS) algorithm $\hat{\boldsymbol{\beta}}_{ML}=\left(\mathbf{X}^{\mathbf{T}}\hat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\hat{\mathbf{W}}\hat{\mathbf{z}}$, where $\hat{\mathbf{W}}=\mathrm{diag}(\hat{\mu}_i)$ and $\hat{\mathbf{z}}$ is a vector in which the i^{th} element is given by

$$\hat{z}_i=\log(\hat{\mu}_i)+\frac{y_i-\hat{\mu}_i}{\hat{\mu}_i}.$$

Under some regularity conditions of MLE, the MLE of $\boldsymbol{\beta}$, has an asymptotic normal distribution, as $n \to \infty$.

2.2. Estimation Strategies

In this study, we consider the three widely-used penalized Poisson regression (PPR). The general formula of PPR is defined as

$$PPR = \ell(\boldsymbol{\beta}) + \lambda P_{\lambda}(\boldsymbol{\beta}), \tag{5}$$

where $P_{\lambda}(\boldsymbol{\beta})$ is the penalty function and λ is a defined tuning parameter (nonnegative regularization parameter: $\lambda > 0$) and it controls the strength of the independent variables. As when λ takes a larger value, more weight will be given to the penalty term.

1) Ridge regression

When the independent variables are a highly correlated matrix of cross products, $\mathbf{X^T\hat{W}X}$ is ill-conditioned. This leads to instability and high covariance in the MLE. In this situation, the MLE loses its efficiency and it becomes very hard to interpret the estimated parameter, since the vector of estimated coefficients is on average too large. However, when multicollinearity occurs in model (3), the ridge regression method can be applied to count data (Månsson et al. 2011). The ridge regression can produce the coefficients that minimize the negative log-likelihood, subject to the L_2 penalty on $\boldsymbol{\beta}$. The ridge regression estimation of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}}_{\lambda}^{\text{Ridge}} = \underset{\beta}{\operatorname{argmin}} \left(-\sum_{i=1}^{n} \left(y_i \mathbf{X_i}^T \boldsymbol{\beta} - \exp(\mathbf{X_i}^T \boldsymbol{\beta}) - \ln y_i ! \right) + \lambda \sum_{j=1}^{p} \beta_i^2 \right),$$

and $L_2 = \sum_{j=1}^{p} \beta_j^2 \le t$. Here, λ is the ridge tuning parameter which controls the shrinkage estimates of $\boldsymbol{\beta}$. Its values can be obtained by using cross validation.

2) LASSO

The application of LASSO to the Poisson regression model was first proposed by Park and Hastie (2007). It is a widely-used technique for simultaneous variable selection and parameter estimation. This method is in some sense similar to ridge regression but more shrinks some coefficients to zero. For moderate values of tuning parameter t in $L_1 = \sum_{j=1}^{p} \beta_j \le t$, many $\hat{\beta}_i$'s go to zero. The use of LASSO is most appropriate when one believes that the effect is sparse, so that the response can be explained by the small number of predictors and the rest of them have no effect. This means that LASSO can be regarded as a type of variable selection method, in which the responding predictor X_i is effectively eliminated from the regression when $\hat{\beta}_i = 0$. In contrast, ridge regression does not eliminate any variables, but simply makes $\hat{\beta}_i$ smaller. It computes the coefficients that minimize the negative log-likelihood, subject to the L_1 penalty on $\boldsymbol{\beta}$. The LASSO estimation of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}}_{\lambda}^{\text{Lasso}} = \underset{\beta}{\operatorname{argmin}} \left(-\sum_{i=1}^{n} \left(y_i \mathbf{X_i}^T \boldsymbol{\beta} - \exp(\mathbf{X_i}^T \boldsymbol{\beta}) - \ln y_i ! \right) + \lambda \sum_{j=1}^{p} \beta_i \right),$$

therefore $L_1 = \sum_{j=1}^{p} \beta_j \le t$ and λ is the tuning parameter.

3) Adaptive LASSO

Adaptive LASSO for a Poisson regression model was proposed by Park and Hastie (2007) based on Fan and Li (2001). It improves the efficiency of parameter estimation and increases variable selection by applying weights. The idea of adaptive LASSO is to give large weights to inactive variables (or variable having no effect), and thus to heavily shrink their associated coefficients. By giving small weights to active variables, it slightly shrinks the corresponding coefficients. It computes the coefficients that minimize the negative log-likelihood subject to the L_1 penalty on the $\boldsymbol{\beta}$. The adaptive LASSO estimation of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}}_{\lambda}^{\text{Adap Lasso}} = \underset{\beta}{\operatorname{argmin}} \left(-\sum_{i=1}^{n} \left(\mathbf{y}_i \mathbf{X}_i' \boldsymbol{\beta} - \exp(\mathbf{X}_i' \boldsymbol{\beta}) - \ln y_i ! \right) + \lambda \sum_{j=1}^{p} |\beta_i| w_i \right),$$

where $L_1 = \sum_{j=1}^{p} \beta_j \le t$ and λ is the tuning parameter. Here w_i is an adaptive weight defined as $w_i = \left| \hat{\beta}_i^* \right|^{-\tau}$ for some positive τ and $\hat{\beta}_i^*$ is a root-n-consistent estimator of β.

2.3. Correlation structure

In this study, we consider the following three correlation structure, which are widely applied in many fields.

1) Constant correlation structure

Our first correlation structure assumes that there is a constant correlation between any two variables. The correlation matrix is shown below. This might be more realistic than simply assuming that all of these entries are zero. Let

$$\Sigma_k = \begin{bmatrix} 1 & \rho & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \rho & \dots & \rho \\ \rho & \rho & 1 & \rho & \dots & \rho \\ \rho & \rho & \rho & 1 & \dots & \rho \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \rho & \dots & 1 \end{bmatrix}_{k \times k},$$

where k denote a positive integer (the number of explanatory variables) and $\rho \in [0, 1]$.

2) Toeplitz correlation structure

This correlation structure assumes that each pair of adjacent variables is highly correlated and the correlations are reduced when the members of the pair are widely separated. The correlations between the i^{th} and j^{th} observations decay exponentially with respect to $|i - j|$. The corresponding correlation structure is

$$\Sigma_k = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \dots & \rho^{k-1} \\ \rho & 1 & \rho & \rho^2 & \dots & \rho^{k-2} \\ \rho^2 & \rho & 1 & \rho & \dots & \rho^{k-3} \\ \rho^3 & \rho^2 & \rho & 1 & \dots & \rho^{k-4} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{k-1} & \rho^{k-2} & \rho^{k-3} & \rho^{k-4} & \dots & 1 \end{bmatrix}_{k\times k},$$

where k denote a positive integer (the number of explanatory variables) and $\rho \in [0,1]$.

3) Hub Toeplitz correlation structure

The hub correlation structure assumes a known correlation between a hub observation (typically the first variable) and each of the other variables. Moreover, it is typically assumed that the correlation between the 1st and the i^{th} observation decays as i increases. As a typical example from the literature, suppose that the first row (and hence column) of a $(g \times g)$ correlation matrix (A) has the prescribed values $A_{11} = 1$, $A_{1i} = \rho_{\max} - (\rho_{\max} - \rho_{\min})\left(\frac{i-2}{g-2}\right)^{\gamma}$, which decreases linearly if $\gamma = 1$ from $A_{12} = \rho_{\max}$ to $A_{1g} = \rho_{\min}$ for $2 \le i \le g$. This model is considered by Zhang et al. (2005) and Langfelder et al. (2007). For the sake of simplicity, we consider the linear case $\gamma = 1$ and adopt a more convenient notation. Rather than specifying $\rho_{\max}$ and $\rho_{\min}$, we specify only $\rho_{\max}$ and work instead with the step size $\tau = (\rho_{\max} - \rho_{\min})/(g-2)$. After specifying the first row, there are a variety of ways to generate the remainder of such a correlation matrix. So, the correlation matrix is

$$\Sigma_k = \begin{bmatrix} 1 & \alpha_{k,2} & \alpha_{k,3} & \alpha_{k,4} & \dots & \alpha_{k,g_k} \\ \alpha_{k,2} & 1 & \alpha_{k,2} & \alpha_{k,3} & \dots & \alpha_{k,g_k-1} \\ \alpha_{k,3} & \alpha_{k,2} & 1 & \alpha_{k,2} & \dots & \alpha_{k,g_k-2} \\ \alpha_{k,4} & \alpha_{k,3} & \alpha_{k,2} & 1 & \dots & \alpha_{k,g_k-3} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha_{k,g_k-1} & \alpha_{k,g_k-1} & \alpha_{k,g_k-2} & \alpha_{k,g_k-3} & \dots & 1 \end{bmatrix}_{k\times k}.$$

3. Simulations

The simulations were carried out to examine the median of mean square error (mPMSE) for ridge regression, LASSO, and adaptive LASSO; the mPMSE could be an indicator of the performance. The variable selections were also compared by using two criteria of incorrect variable selection: False Negative Rate (FNR) and False Positive Rate (FPR) defined as

$$\text{FNR} = \left\{ j:\ \beta_j \neq 0 \text{ but } \hat{\beta}_j = 0 \right\} \text{ and } \text{FPR} = \left\{ j:\ \beta_j = 0 \text{ but } \hat{\beta}_j \neq 0 \right\}.$$

FNR and FPR are so-called Identify Criterion 1 (IC1) and Identify Criterion 2 (IC2), respectively. In simulations, sample sizes (n) are 25 and 50, and the number of independent variables (p) is set to be 50, 100, and 200. Each configuration was run 5,000 times to get a stable result, implemented in R program (R Core Team 2015). Four cases are created as following: 1) For Cases 1 and 2, the independent variables were divided into two groups. The first group contains only independent variables whose $\beta_j \neq 0$ and the corresponding variables are so-called "active". Among these

variables, the correlation value is set to be 0.5 or 0.9. The second group has only independent variables whose $\beta_j = 0$ and they will be called "inactive" variables. The same correlation is given to this group; 2) For Cases 3 and 4, the active and inactive independent variables are combined to 1 group and the correlation value is set to be 0.5 or 0.9. All cases depended on three correlation models: the constant, Toeplitz, and hub Toeplitz correlation models.

4. Results

The results will be divided into 3 cases according to the correlation models. We noted that ridge regression does not perform the variable selection. Its FNR and FPR results then did not present.

4.1. Constant correlation model

The mPMSE values for different configurations of n and p are summarized in Table 1 and graphically represented in Figures 1 and 2 for aid comparison. As can be seen, the adaptive LASSO tended to the lowest mPMSE followed by LASSO and ridge regression, respectively. When r increased, the performance of all estimators improved, as their mPMSE decreased. For fixed p, the mPMSEs of all estimators increased as n increased in 1 group cases, but not in 2 group cases. Considering Cases 1, 2, or 3 (except Case 4) with r of 0.5 or 0.9, the highest mPMSE occurred when $n = 50,\ p = 50$.

Table 1 Median of prediction mean square error of three methods in constant correlation model

n	p	2 groups			1 group		
		Ridge	LASSO	Adaptive LASSO	Ridge	LASSO	Adaptive LASSO
		$r = 0.5$					
		Case 1			Case 3		
25	50	9.193113	7.315495	6.273284*	10.65240	6.189153	5.076029*
25	100	14.87306	10.01593	8.624981*	7.086174	6.230699	4.965397*
25	200	10.44676	9.609530	6.869651*	11.21742	10.12269	7.63458*
50	50	17.00520	10.86592	8.985308*	23.89061	14.69172	13.18459*
50	100	10.63253	8.413216	6.737952*	10.47410	8.080800	6.840008*
50	200	11.62968	9.861743	7.209156*	13.36298	10.86931	8.42348*
		$r = 0.9$					
		Case 2			Case 4		
25	50	3.516828	2.832158	2.606889*	4.59058	3.95301	3.76701*
25	100	5.072896	4.570755	4.219117*	4.64637	3.95966	3.78637*
25	200	4.420250	3.665096	3.500912*	4.33225	3.65982	3.53580*
50	50	7.983044	6.146937	5.742676*	7.18089	5.97884	6.07353*
50	100	7.369058	6.107473	5.450372*	9.28073	7.48053	7.17576*
50	200	4.897087	3.461501	3.307722*	6.66773	5.65949	5.26692*

*The lowest value of median of prediction mean square error

The FNR and FPR values of LASSO and adaptive LASSO results are reported in Table 2. Overall, the FNR of both methods increased as p increased for fixed n, in contrast to the FPR. The FNR and FPR behaviors were similar when n increases for fixed p. In almost all situations, LASSO performed better in terms of FNR, indicating that adaptive LASSO eliminated too many significant predictors. However, LASSO also kept too many noises in the resulting model, in which it had higher FPR.

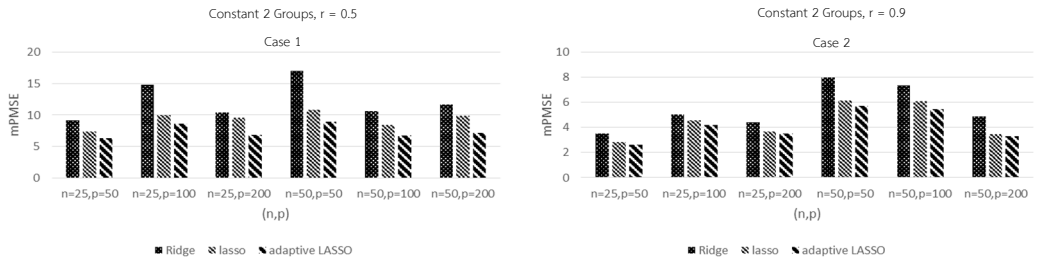

Figure 1 Comparisons of mPMSE for constant correlation model for 2 groups, $r = 0.5$ and 0.9

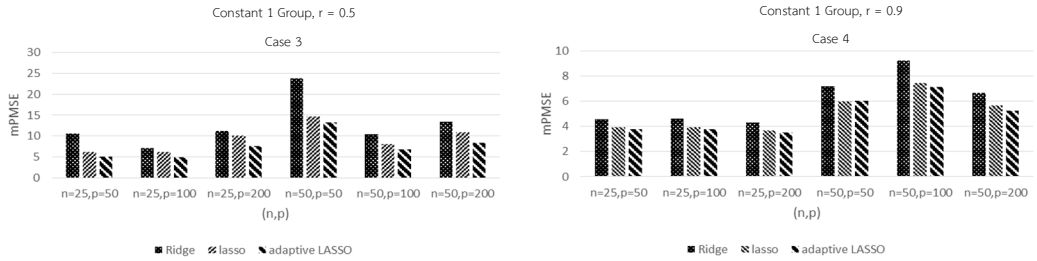

Figure 2 Comparisons of mPMSE for constant correlation model for 1 group, $r = 0.5$ and 0.9

Table 2 Probability of incorrect selection by LASSO and adaptive LASSO

n	p	Independent variables divided into 2 groups				Independent variables not divided (1 group)			
		LASSO		Adaptive LASSO		LASSO		Adaptive LASSO	
		FNR	FPR	FNR	FPR	FNR	FPR	FNR	FPR
		$r = 0.5$							
		Case 1				Case 3			
25	50	0.1093	0.4000	0.1467	0.3920	0.1717	0.3913	0.1892	0.3886
25	100	0.1640	0.1652	0.1693	0.1611	0.1759	0.1699	0.2521	0.1679
25	200	0.1887	0.0779	0.3013	0.0764	0.2327	0.0792	0.3030	0.0789
50	50	0.2453	0.3549	0.2927	0.3566	0.3049	0.3441	0.2561	0.3413
50	100	0.3133	0.1587	0.3060	0.1514	0.3095	0.1600	0.3780	0.1546
50	200	0.3207	0.0760	0.4673	0.0744	0.3321	0.0771	0.4984	0.0747
		$r = 0.9$							
		Case 2				Case 4			
25	50	0.0573	0.3911	0.0973	0.3909	0.1071	0.4090	0.1002	0.4087
25	100	0.0720	0.1652	0.1167	0.1640	0.1451	0.1719	0.1465	0.1719
25	200	0.1413	0.0759	0.1720	0.0750	0.1631	0.0793	0.1773	0.0792
50	50	0.0707	0.3734	0.1267	0.3694	0.1679	0.3837	0.1235	0.3930
50	100	0.1440	0.1578	0.1507	0.1519	0.2599	0.1653	0.2437	0.1658
50	200	0.1933	0.0732	0.2507	0.0700	0.2953	0.0786	0.3083	0.0785

4.2. Toeplitz correlation model

The mPMSE values for different configurations of n and p are summarized in Table 3 and graphically represented in Figures 3 and 4.

Table 3 Median of prediction mean square error of three methods in Toeplitz correlation model

n	p	Independent variables divided into 2 groups			Independent variables not divided (1 group)		
		Ridge	LASSO	Adaptive LASSO	Ridge	LASSO	Adaptive LASSO
		$r = 0.5$					
		Case 1			Case 3		
25	50	21.495140	14.501810	12.049180*	16.153980	11.829430	8.670539*
25	100	7.173000	6.938153	5.175408*	5.411574	4.330981	3.507906*
25	200	8.346166	9.250106	5.660197*	8.194067	8.516340	5.741609*
50	50	15.700180	10.457520	9.403252*	29.174740	20.41302	14.77027*
50	100	16.843000	10.795610	8.288666*	21.060200	15.97994	12.22886*
50	200	19.179300	16.364800	10.478430*	20.860800	17.23301	12.59359*
		$r = 0.9$					
		Case 2			Case 4		
25	50	4.524129	4.248000	3.423570*	4.303445	3.923764	3.483195*
25	100	3.201618	3.007115	2.431623*	4.786392	4.681250	3.264642*
25	200	5.307936	4.837770	4.139178*	5.655110	5.745000	4.407430*
50	50	6.229060	5.248282	4.718587*	5.583285	5.517457	4.928481*
50	100	6.146114	5.650000	4.329029*	6.607350	6.140578	5.245554*
50	200	6.549703	6.637454	5.035816*	7.230346	6.871357	5.275375*

*The lowest value of median of prediction mean square error

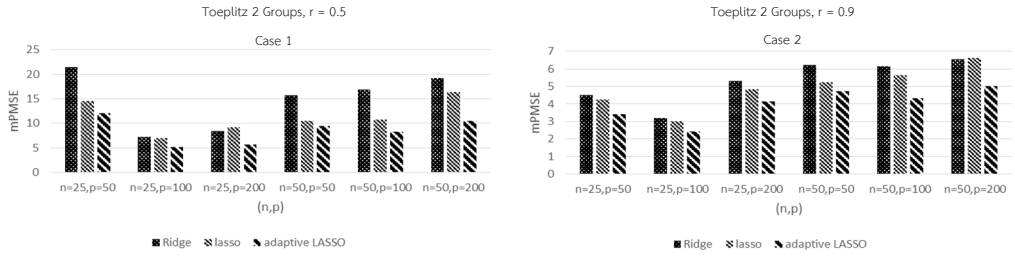

Figure 3 Comparisons of mPMSE for Toeplitz correlation model, independent variable 2 groups, $r = 0.5$ and 0.9

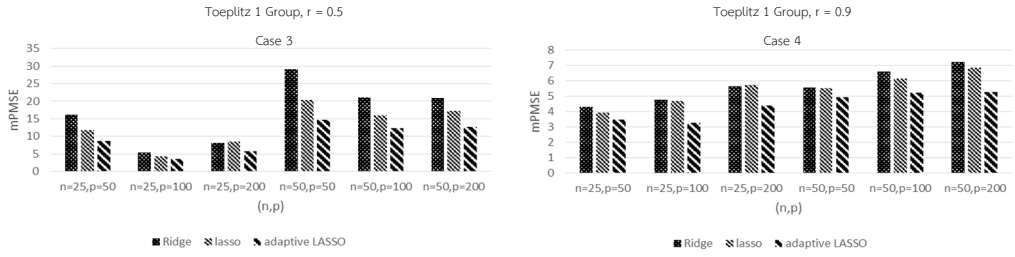

Figure 4 Comparisons of mPMSE for Toeplitz correlation model, independent variable 1 group, $r = 0.5$ and 0.9

The main findings were as follows:

(i) The adaptive LASSO had the highest performance among all others across all cases.

(ii) LASSO outperformed Ridge regression in the wider range of parameters.

(iii) The mPMSE reduction of all estimators increased when r became stronger for both 1 and 2 groups.

(iv) For fixed p, the performance of all estimators decreased as n increased in all cases, except Case 1.

The variable selection performance of LASSO and adaptive LASSO in terms of FNR and FPR are reported in Table 4. The LASSO, having the smaller FNR, had stronger performance in selecting active predictors, but adaptive LASSO, having the smaller FPR, performed better in removing the inactive predictors from the model. Both methods significantly improved the performance in eliminating the noises when either p or n increased, however their performance in selecting the significant predictors also deteriorated, as FPR increased, whereas FPR decreased. These results are similar those from constant correlation model.

4.3. Hub Toeplitz correlation model

The mPMSE values for the hub Toeplitz correlation model are summarized in Table 5 and graphically represented in Figures 5 and 6.

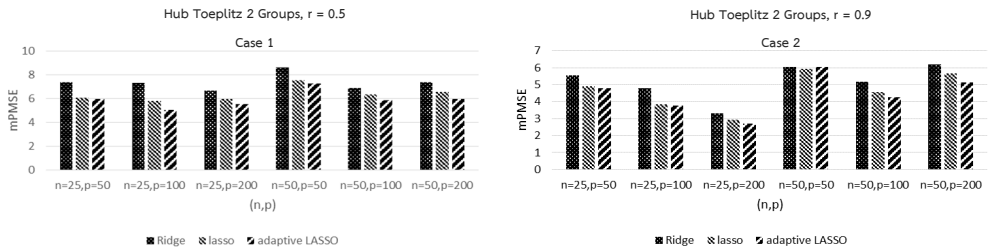

Figure 5 Comparisons of mPMSE for hub Toeplitz correlation model, independent variable 2 groups, r = 0.5 and 0.9

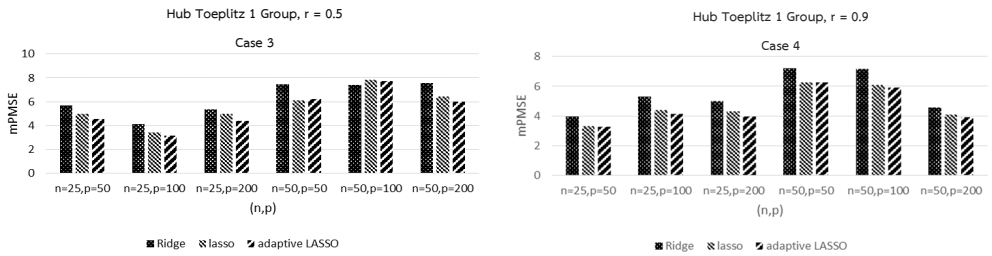

Figure 6 Comparisons of mPMSE for hub Toeplitz correlation model, independent variable 1 group, r = 0.5 and 0.9

Table 5 Median of prediction mean square error of three methods in hub Toeplitz correlation model

n	p	Independent variables divided into 2 groups			Independent variables not divided (1 group)		
		Ridge	LASSO	Adaptive LASSO	Ridge	LASSO	Adaptive LASSO
		$r=0.5$					
		Case 1			Case 3		
25	50	7.398150	6.111601	5.959504*	5.663284	4.969779	4.558758*
25	100	7.360890	5.836827	5.079702*	4.126709	3.416840	3.144514*
25	200	6.683900	5.975250	5.553250*	5.339324	5.004806	4.378678*
50	50	8.662850	7.574599	7.282740*	7.452506	6.105046	6.224579*
50	100	6.915100	6.382655	5.860514*	7.397727	7.853712	7.709643*
50	200	7.392311	6.585958	5.960586*	7.572047	6.458922	5.999585*
		$r=0.9$					
		Case 2			Case 4		
25	50	5.545045	4.898264	4.792709*	3.950863	3.293000	3.276790*
25	100	4.794218	3.864274	3.781592*	5.309332	4.410250	4.140341*
25	200	3.303360	2.941000	2.723301*	5.010201	4.326250	3.963443*
50	50	6.033837	5.919316	6.033541*	7.220565	6.260250	6.280313*
50	100	5.182348	4.570781	4.278688*	7.167214	6.072605	5.916763*
50	200	6.183978	5.669062	5.151912*	4.577000	4.076177	3.898372*

*The lowest value of median of prediction mean square error

The FNR and FPR values of LASSO and Adaptive LASSO are reported in Table 6.

Table 6 Probability of incorrect selection by LASSO and adaptive LASSO

n	p	Independent variables divided into 2 groups			Independent variables not divided (1 group)		
		Ridge	LASSO	Adaptive LASSO	Ridge	LASSO	Adaptive LASSO
		$r=0.5$					
		Case 1			Case 3		
25	50	0.0557	0.3928	0.0721	0.3871	0.0902	0.4053
25	100	0.1283	0.1639	0.1375	0.1615	0.1218	0.1690
25	200	0.1295	0.0778	0.1583	0.0768	0.1426	0.0798
50	50	0.1350	0.3800	0.1489	0.3776	0.1733	0.3874
50	100	0.1531	0.1623	0.1555	0.1583	0.1948	0.1647
50	200	0.2016	0.0743	0.1991	0.0723	0.2603	0.0778
		$r=0.9$					
		Case 2			Case 4		
25	50	0.0777	0.3969	0.0618	0.3921	0.1201	0.4047
25	100	0.1083	0.1651	0.1049	0.1631	0.1364	0.1707
25	200	0.1215	0.0774	0.1280	0.0758	0.1590	0.0797
50	50	0.1544	0.3761	0.1010	0.3742	0.1132	0.3913
50	100	0.1810	0.1602	0.1619	0.1559	0.2305	0.1686
50	200	0.1936	0.0739	0.1921	0.0709	0.2585	0.0772

The mPMSE results and the variable selection results for the hub Toeplitz correlation model were similar to those from previous models, and are not reported here.

5. Real Data Example

We next test the three methods using real data: These data were taken from a 2016 assessment of software engineering team working in an educational setting. They were used to predict the number of student teams based on observations. The dataset included observations of 64 teams and 79 explanatory variables. This was split into a training set of 51 observations and a test set of 13 observations. Model fitting and tuning parameter setting were done using 5-folds cross validation in the training set.

When analyzed using mPMSE reported in Table 7, adaptive LASSO performed best in terms of prediction error, followed by LASSO and then ridge regression. From the FNR and FPR results reported in Table 8, LASSO produced FPR values equal to adaptive LASSO. However, LASSO showed higher performance in selecting the significant predictors.

Table 7 Median of prediction mean square error of three methods

Ridge	LASSO	Adaptive LASSO
2.345	2.241	2.217

Table 8 Probability of incorrect selection by LASSO and adaptive LASSO

LASSO		Adaptive LASSO	
FNR	FPR	FNR	FPR
0.333	0.187	0.733	0.187

6. Conclusions

We compared the performance of ridge regression, LASSO and adaptive LASSO in Poisson regression for high-dimensional sparse data with multicollinearity. In both simulations and tests using real data, adaptive LASSO was demonstrated to outperform the other two methods in terms of the mean squares error. Adaptive LASSO and LASSO performed similarly in terms of incorrect variable selection. Ridge regression was shown to have the best performance in the presence of multicollinearity between variables in the Poisson regression model. We conclude that Ridge regression performs the best in the hub Toeplitz correlation model, followed by the constant correlation model, and then the Toeplitz correlation model. Taking account of both prediction and the probability of incorrect variable selection, adaptive LASSO was shown to have the best performance when analyzing high-dimensional sparse data.

Acknowledgments

The authors gratefully acknowledge the financial support provided by Thammasat University. We would also like to thank the editor and referees for their valuable suggestions on the revision of this paper.

References

Ahmed SE, Yüzbaşı B. Big data analytics: integrating penalty strategies. Int J Ind Eng Manag Sci. 2016; 11(2): 105-115.

Algamal ZY, Lee MH. Adjusted adaptive LASSO in high-dimensional Poisson regression model. Mod Appl Sci. 2015; 9(4): 170-177.

Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc. 2001; 96(456): 1348-1360.

Gao X, Ahmed SE, Feng Y. Post selection shrinkage estimation for high-dimensional data analysis. Appl Stoch Models Bus Ind. 2017; 33(2): 97-120.

Hardin J, Garcia SR, Golan D. A method for generating realistic correlation matrices. Ann Appl Stat. 2013; 7(13): 1733-1762.

Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970; 12(1): 55-67.

Hossain S, Ahmed SE. Shrinkage and penalty estimators of a Poisson regression model. Aust NZ J Stat. 2012; 54(3): 359-373.

Ivanoff S, Picard F, Rivoirard V. Adaptive LASSO and group-LASSO for functional Poisson regression. J Mach Learn Res. 2016; 17(55): 1-46.

Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for R. Bioinformatics. 2007; 24(5): 719-720.

Månsson K, Shukur G. A Poisson ridge regression estimator. Econ Model. 2011; 28(4): 1475-1481.

Oyeyemi GM, Ogunjobi EO, Folorunsho AI. On performance of shrinkage methods-a Monte Carlo study. Int J Stat Appl. 2015; 5(2): 72-76.

Park MY, Hastie T. L1-regularization path algorithm for generalized linear models. J R Stat Soc Series B Stat Methodol. 2007; 69(4): 659-677.

Pungpapong V. A Brief review on high-dimensional linear regression. Scholarly Article of Statistical, Chulalongkorn University, Bangkok. 2014.

R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2015.

Yüzbası B, Ahmed SE, Güngör M. Improved penalty strategies in linear regression models. REVSTAT-Stat J. 2017a; 15(2): 251-276.

Yüzbaşı B, Arashi M, Ahmed SE. Shrinkage estimation strategies in generalized ridge regression models under low/high-dimension regime. arXiv preprint arXiv:1707.02331. 2017b.

Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Gen Mol Biol. 2005; 4(1): Article17. doi:10.2202/1544-6115.1128.

Zou H. The adaptive LASSO and its oracle properties. J Am Stat Assoc. 2006; 101(476): 1418-1429.