Computational Statistics and Data Analysis 52 (2008) 5277–5286

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A note on adaptive group lasso

Hansheng Wang *, Chenlei Leng

Peking University, China
National University of Singapore, Singapore

ARTICLE INFO

Article history:
Received 30 January 2007
Received in revised form 28 April 2008
Accepted 10 May 2008
Available online 20 May 2008

ABSTRACT

Group lasso is a natural extension of lasso and selects variables in a grouped manner. However, group lasso suffers from estimation inefficiency and selection inconsistency. To remedy these problems, we propose the adaptive group lasso method. We show theoretically that the new method is able to identify the true model consistently, and the resulting estimator can be as efficient as *oracle*. Numerical studies confirmed our theoretical findings.

1. Introduction

Since its first proposal by Tibshirani (1996), the *least absolute shrinkage and selection operator* (lasso) has generated much interest in the statistical literature (Fu, 1998; Knight and Fu, 2000; Fan and Li, 2001; Efron et al., 2004). The key strength of lasso lies in its ability to do simultaneous parameter estimation and variable selection. However, recent research suggests that the traditional lasso estimator may not be fully efficient (Fan and Li, 2001), and its model selection result could be inconsistent (Leng et al., 2006; Yuan and Lin, 2007; Zou, 2006). The major reason accounting for such a deficiency is that lasso applies the same amount of shrinkage for each regression coefficient. As a simple solution, Zou (2006) modified the lasso penalty so that different amounts of shrinkage are allowed for different regression coefficients. Such a modified lasso method was referred to as adaptive lasso (Zou, 2006, aLasso). It has been shown theoretically that the aLasso estimator is able to identify the true model consistently, and the resulting estimator is as efficient as *oracle*. Similar methods were also developed for Cox's proportional hazard model (Zhang and Lu, 2007), least absolute deviation regression (Wang et al., 2007a), and linear regression with autoregressive residuals (Wang et al., 2007b).

Both the original lasso and adaptive lasso were designed to select variables individually. However, there are situations where it is desirable to choose predictive variables in a grouped manner. The multifactor analysis of variance model is a typical example. To this end, Yuan and Lin (2006) developed the group lasso (gLasso) method, which penalizes the grouped coefficients in a similar manner to lasso. Hence, it is expected that gLasso in Yuan and Lin (2006) suffers from the estimation inefficiency and selection inconsistency in the same way as lasso. As a remedy, we propose the adaptive group lasso (agLasso) method. It is similar to adaptive lasso but has the capability to select variables in a grouped manner. We show theoretically that the proposed agLasso estimator is able to identify the true model consistently, and the resulting estimator is as efficient as *oracle*. Numerical studies confirmed our theoretical findings.

The rest of the article is organized as follows. The agLasso method is proposed in the next section, and its theoretical properties are established in Section 3. Simulation results are reported in Section 4 and one real dataset is analyzed in Section 5. All technical details are presented in the Appendix.

* Corresponding address: Peking University, Guanghua School of Management, 100871 Beijing, China. Tel.: +86 10 6275 7915.
E-mail address: hansheng@gsm.pku.edu.cn (H. Wang).

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.05.006

2. Adaptive group lasso

2.1. Model and notations

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a total of n independent and identically distributed random vectors, where $y_i \in \mathcal{R}^1$ is the response of interest and $x_i \in \mathcal{R}^d$ is the associated d-dimensional predictor. Furthermore, it is assumed that x_i can be grouped into p factors as $x_i = (x_{i1}^\top, \ldots, x_{ip}^\top)^\top$, where $x_{ij} = (x_{ij1}, \ldots, x_{ijd_j})^\top \in \mathcal{R}^{d_j}$ is a group of d_j variables. In such a situation, it is practically more meaningful to identify important factors instead of individual variables (Yuan and Lin, 2006). In this paper, we use the terms factor and group interchangeably to indicate the grouping of variables. For example, a categorical variable may be represented by a few indicator variables and these indicator variables form a factor. On the other hand, we may use a few polynomials to represent the effect of a continuous variable and group these polynomials as a single factor. In order to model the dependence relationship between the responses y_i and x_i, the following typical linear regression model is assumed

$$y_i = \sum_{j=1}^{p} x_{ij}^\top \beta_j + e_i = x_i^\top \beta + e_i,$$

where $\beta_j = (\beta_{j1}, \ldots, \beta_{jd_j})^\top \in \mathcal{R}^{d_j}$ is the regression coefficient vector associated with the jth factor and β is defined to be $\beta = (\beta_1^\top, \ldots, \beta_p^\top)^\top$. Without loss of generality, we assume that only the first $p_0 \le p$ factors are relevant. That is, we assume that $\|\beta_j\| \neq 0$ for $j \le p_0$ and $\|\beta_j\| = 0$ for $j > p_0$, where $\|\cdot\|$ is the L_2 norm of a vector.

2.2. The agLasso estimator

For simultaneous parameter estimation and factor selection, Yuan and Lin (2006) proposed the following penalized least squares type objective function with the group lasso (gLasso) penalty

$$\sum_{i=1}^{n} \frac{1}{2} \left(y_i - \sum_{j=1}^{p} x_{ij}^\top \beta_j \right)^2 + n\lambda \sum_{j=1}^{p} \|\beta_j\|.$$

Note that if the number of variables contained in each factor is indeed one (i.e., $d_j = 1$), the above gLasso objective function reduces to the usual lasso. However, if there do exist some factors containing more than one variable, the above gLasso estimator has the capability to select those variables in a grouped manner.

As one can be seen, gLasso penalizes each factor in a very similar manner as the usual lasso. In other words, same tuning parameter λ is used for each factor without assessing their relative importance. In a typical linear regression setting, it has been shown that such an excessive penalty applied to the relevant variables can degrade the estimation efficiency (Fan and Li, 2001) and affect the selection consistency (Leng et al., 2006; Yuan and Lin, 2007; Zou, 2006). Therefore, we can reasonably expect that gLasso suffers the same drawback. To overcome such a limitation, we borrow the adaptive lasso idea and propose the following adaptive group lasso (agLasso)

$$Q(\beta) = \sum_{i=1}^{n} \frac{1}{2} \left(y_i - \sum_{j=1}^{p} x_{ij}^\top \beta_j \right)^2 + n \sum_{j=1}^{p} \lambda_j \|\beta_j\|. \tag{2.1}$$

Then, minimizing the above objective function produces the agLasso estimator $\hat{\beta}$. As can be seen, the key difference between the agLasso and gLasso is that the agLasso allows for different tuning parameters used for different factors. Such a flexibility in turn produces different amounts of shrinkage for different factors. Intuitively, if a relatively larger amount of shrinkage is applied to the zero coefficients and a relatively smaller amount is used for the nonzero coefficients, an estimator with a better efficiency can be obtained. For practical implementation, one usually does not know which factor is important and which one is not. However, without such prior knowledge, simple estimators of λ_j can be obtained in a similar manner to Zou (2006). With those estimated tuning parameters, we are able to show theoretically that the proposed agLasso estimator can indeed identify the true model consistently and the resulting estimator is as efficient as *oracle*.

2.3. Tuning parameter selection

For practical implementation, one has to decide the values of the tuning parameters (i.e, λ_j). Traditionally, cross-validation (CV) or generalized cross-validation (GCV) have been widely used. However, those computationally intensive methods can hardly be useful for agLasso, simply because there are too many tuning parameters. As a simple solution (Zou, 2006; Wang et al., 2007b; Zhang and Lu, 2007), we consider

$$\lambda_j = \lambda \|\tilde{\beta}_j\|^{-\gamma}, \tag{2.2}$$

where $\tilde{\beta} = (\tilde{\beta}_1^\top, \ldots, \tilde{\beta}_p^\top)^\top$ is the unpenalized least squares estimator and $\gamma > 0$ is some pre-specified positive number. For example, $\gamma = 1$ is used for our simulation study and real data analysis. Then, the original p-dimensional tuning parameter selection problem for $(\lambda_1, \ldots, \lambda_p)$ reduces to a univariate problem for λ only. Thereafter, any appropriate selection method can be used. In our numerical studies, the following selection criteria were considered:

$$\begin{aligned}
\text{Cp} &= \frac{\|Y - X\hat{\beta}\|^2}{\tilde{\sigma}^2} - n + 2df \\
\text{GCV} &= \frac{\|Y - X\hat{\beta}\|^2}{(1 - n^{-1} \times df)^2} \\
\text{AIC} &= \log\left(\frac{1}{n}\|Y - X\hat{\beta}\|^2\right) + 2df/n \\
\text{BIC} &= \log\left(\frac{1}{n}\|Y - X\hat{\beta}\|^2\right) + \log n \times df/n.
\end{aligned} \tag{2.3}$$

Note that df is the associated degree of freedom as defined in Yuan and Lin (2006), given by

$$df = \sum_{j=1}^{p} I\left\{\|\hat{\beta}_j\| > 0\right\} + \sum_{j=1}^{p} \frac{\|\hat{\beta}_j\|}{\|\tilde{\beta}_j\|}(d_j - 1),$$

where $\tilde{\sigma}^2 = \|Y - X\tilde{\beta}\|^2/(n - df)$ is the usual variance estimator associated with $\tilde{\beta}$.

3. Theoretical properties

The main theoretical properties of the proposed agLasso estimator are established in this section. For the purpose of easy discussion, we define $a_n = \max\{\lambda_j, j \le p_0\}$ and $b_n = \min\{\lambda_j, j > p_0\}$.

Theorem 1 (*Estimation Consistency*). *If* $\sqrt{n}a_n \to_p 0$, *then* $\hat{\beta} - \beta = O_p(n^{-1/2})$.

Note that "$\to_p$" denotes convergence in probability. From Theorem 1 we know that, as long as the maximal amount of the shrinkage applied to the relevant variables is sufficiently small, $\sqrt{n}$-consistency is assured. Next, we establish the consistency of the agLasso estimator as a variable selection method. To facilitate discussion, some notations need to be defined. Let $\beta_a = (\beta_1^\top, \ldots, \beta_{p_0}^\top)^\top$ to be the vector containing all the relevant factors, and let $\beta_b = (\beta_{p_0+1}^\top, \ldots, \beta_p^\top)^\top$ be the vector containing all the irrelevant factors. Furthermore, let $\hat{\beta}_a$ and $\hat{\beta}_b$ be their associated agLasso estimators. If one knows the true model, the *oracle* estimator can be obtained, which is denoted by $\tilde{\beta}_a$. Standard linear model theory implies that $\sqrt{n}(\tilde{\beta}_a - \beta_a) \to_d N(0, \Sigma_a)$ where Σ_a is the $(d_0 = \sum_{j=1}^{p_0} d_j)$-dimensional covariance matrix of the first p_0 relevant factors.

Theorem 2 (*Selection Consistency*). *If* $\sqrt{n}a_n \to_p 0$ *and* $\sqrt{n}b_n \to_p \infty$, *then* $P(\hat{\beta}_b = 0) \to 1$.

According to Theorem 2, we know that, with probability tending to one, all the zero coefficients must be estimated exactly as 0. On the other hand, by Theorem 1, we know that the estimates for the nonzero coefficients must be consistent. Such a consistency implies that, with probability tending to one, all the relevant variables must be identified with nonzero coefficients. Both Theorems 1 and 2 imply that agLasso does have the ability to identify the true model consistently.

Theorem 3 (*Oracle Property*). *If* $\sqrt{n}a_n \to_p 0$ *and* $\sqrt{n}b_n \to_p \infty$, *then* $\sqrt{n}(\hat{\beta}_a - \beta_a) \to_d N(0, \Sigma_a)$.

Note that "$\to_d$" denotes convergence in distribution. By Theorem 3, we know that, as long as the conditions $\sqrt{n}a_n \to_p 0$ and $\sqrt{n}b_n \to_p \infty$ are satisfied, the resulting estimator is as efficient as *oracle*. For the tuning parameters specified by (2.2), the conditions of Theorem 3 require that

$$\frac{\sqrt{n}\lambda}{\|\tilde{\beta}_j\|^\gamma} \to_p \begin{cases} 0 & j \le p_0 \\ \infty & j > p_0. \end{cases}$$

Now, since $\tilde{\beta}$ is $\sqrt{n}$-consistent, we have that $\|\tilde{\beta}_j\| \to_p \|\beta_j\|$ for $j \le p_0$ and $\|\tilde{\beta}_j\| = O_p(1/\sqrt{n})$ for $j > p_0$. Therefore, as long as $n^{1/2}\lambda \to 0$ and $n^{(1+\gamma)/2}\lambda \to \infty$, the conditions $\sqrt{n}a_n \to_p 0$ and $\sqrt{n}b_n \to_p \infty$ are satisfied. If we write $\lambda = n^\alpha$, then we can take $\alpha \in (-(1+\gamma)/2, -1/2)$ which satisfies the conditions in Theorem 3.

Sparse-group-Lasso-with sparsity.

group-Lasso => group-variable 전부날림.

Table 1
Model selection comparison for teaching evaluation data

Selection method	No. of factors selected			Outsample MSE ($\times 10^{-1}$)		
	agLasso	aLasso	gLasso	agLasso	aLasso	gLasso
Cp	4.51 (0.05)	4.55 (0.05)	5.89 (0.05)	1.917 (0.04)	1.937 (0.04)	1.932 (0.04)
GCV	4.51 (0.05)	4.55 (0.05)	5.89 (0.05)	1.917 (0.04)	1.937 (0.04)	1.932 (0.05)
AIC	4.51 (0.05)	4.57 (0.05)	5.90 (0.05)	1.917 (0.04)	1.936 (0.04)	1.930 (0.05)
BIC	4.06 (0.06)	4.07 (0.03)	4.95 (0.06)	1.936 (0.05)	1.958 (0.04)	1.956 (0.05)

The standard errors are presented in parentheses.

4. Simulation study

group-variable! 쉽게 말하면 범주형 변수 coding

Simulation studies were conducted to evaluate the finite sample performance of agLasso. For comparison purpose, the performance of both aLasso and gLasso were also evaluated. For each simulated dataset, various selection criteria defined in (2.3) were tested. All the examples reported in this section were borrowed from Yuan and Lin (2006).

Example 1. In this example, 15 latent variables $Z_1, \ldots, Z_{15}$ were generated according to a zero mean multivariate normal distribution, whose covariance between Z_i and Z_j is fixed to be $0.5^{|i-j|}$. Subsequently, Z_i is trichotomized as 0, 1, or 2 if it is smaller than $\Phi^{-1}(1/3)$, larger than $\Phi^{-1}(2/3)$, or in between. Then, response Y is generated from

$$Y = -1.2\, I(Z_1 = 0) + 1.8\, I(Z_1 = 1) + 0.5\, I(Z_3 = 0) + I(Z_3 = 1) + I(Z_5 = 0) + I(Z_5 = 1) + \epsilon,$$

where $I(\cdot)$ is the indicator function and the residual ϵ is normally distributed with mean 0 and standard deviation σ. In this example, we have 15 factors (groups), each with 3 dummy variables. The factors associated with Z_1, Z_3 and Z_5 are nonzero. For a relatively complete evaluation, various sample sizes (i.e., $n = 50, 100, 150, 200, 250$) and various noise levels (i.e., $\sigma = 0.5, 1.0, 2.0$) were tested. For each parameter setting, 200 datasets were simulated and the *median relative model error* (MRME) is summarized (Fan and Li, 2001). For each selection method, the percentage of the 200 simulated datasets, at which the true model is correctly identified, is computed. Lastly, the average model size (i.e., the number of factors) was compared. Due to the fact that the simulation results for GCV, Cp, and AIC are very similar, only the results for Cp and BIC are presented in Figs. 1 and 2, respectively. We find that agLasso clearly stands out to be the best estimator for every performance measure, almost every sample size, and every selection criterion. In terms of variable selection, we see that BIC is superior to Cp. The reason is that when there exists a true model, AIC type of criteria (including GCV and Cp) tend to overestimate the model size (Leng et al., 2006; Wang et al., 2007b,c). Subsequently, estimation accuracy using Cp may suffer. A theoretical justification which shows GCV overfits for smoothly clipped absolute deviation (Fan and Li, 2001, SCAD) method is given by Wang et al. (2007c). The same arguments apply to the Cp and agLasso methods too.

Example 2. In this example, we generated 20 covariates $X_1, \ldots, X_{20}$ in the same fashion as in Example 1. However, only the last 10 covariates $X_{11}, \ldots, X_{20}$ were trichotomized in the same manner as described in the first example. The true regression model is fixed to be

$$Y = X_3 + X_3^2 + X_3^3 + \frac{2}{3}X_6 - X_6^2 + \frac{1}{3}X_6^3 + 2\, I(X_{11} = 0) + I(X_{11} = 1) + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$. For this example, there are 20 factors, each associated with one X variable. And among them, there are 3 nonzero factors, associated with X_3, X_6 and X_{11} respectively. Then, the three competing methods (i.e., aLasso, gLasso, agLasso) were compared at different sample sizes (i.e., $n = 100, 200, 300, 400$) and different noise levels (i.e., $\sigma = 0.5, 1.0, 2.0$). The results were summarized in Figs. 3 and 4. The results are very similar to that of Example 1.

5. The teaching evaluation data

In order to demonstrate the usefulness of agLasso in real situation, we present in this section one real example. The data is about the teaching evaluation scores collected from a total of 340 courses taught in Peking University. For each observation, the response of interest is the teaching evaluation score for one particular course, taught in Peking University during the period of 2002–2004. There is only 1 continuous predictor, which is the log-transformed class size (i.e., how many students enrolled in the class). Due to the suspicion of some nonlinear relationship, a third-order polynomial is used to fully characterize the class size effect. In addition to that, there are 5 different categorical variables, which are suspected to have explanatory power for the response. They are, respectively, the instructor's title (assistant professor, associate professor, and full professor), the instructor's gender (male or female), the student type (MBA, Undergraduate, and Graduate), the semester (Spring or Fall), and the year (2002, 2003, and 2004). For a fair evaluation, we randomly split the 340 observations into two parts. One part contains a total of 300 observations, which are used to build the model. The other part contains the remaining 40 observations, which are used to evaluate the outsample forecasting error. For a reliable comparison, we repeated such a procedure a total of 100 times with the key findings reported in Table 1. As one can see, regardless of which selection

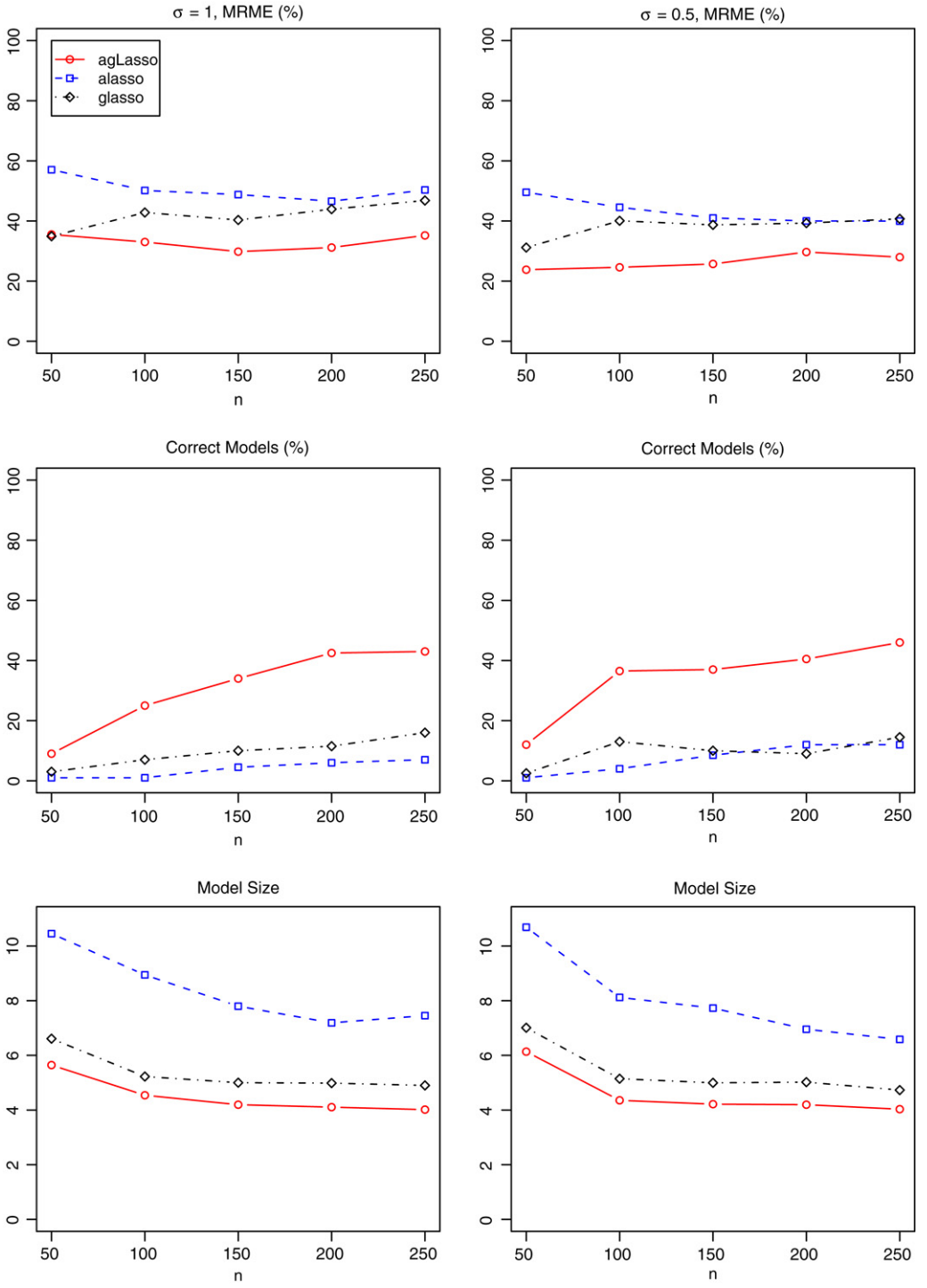

Fig. 1. Model 1 with Cp as the selection criterion.

method is used, the optimal model selected by agLasso consistently demonstrated the smallest average model size and the best prediction accuracy.

Acknowledgements

We are grateful to the two referees, the associate editor, and the editor for their helpful comments. Wang's research is supported in part by NSFC (10771006). Leng's research is supported in part by NUS research grants.

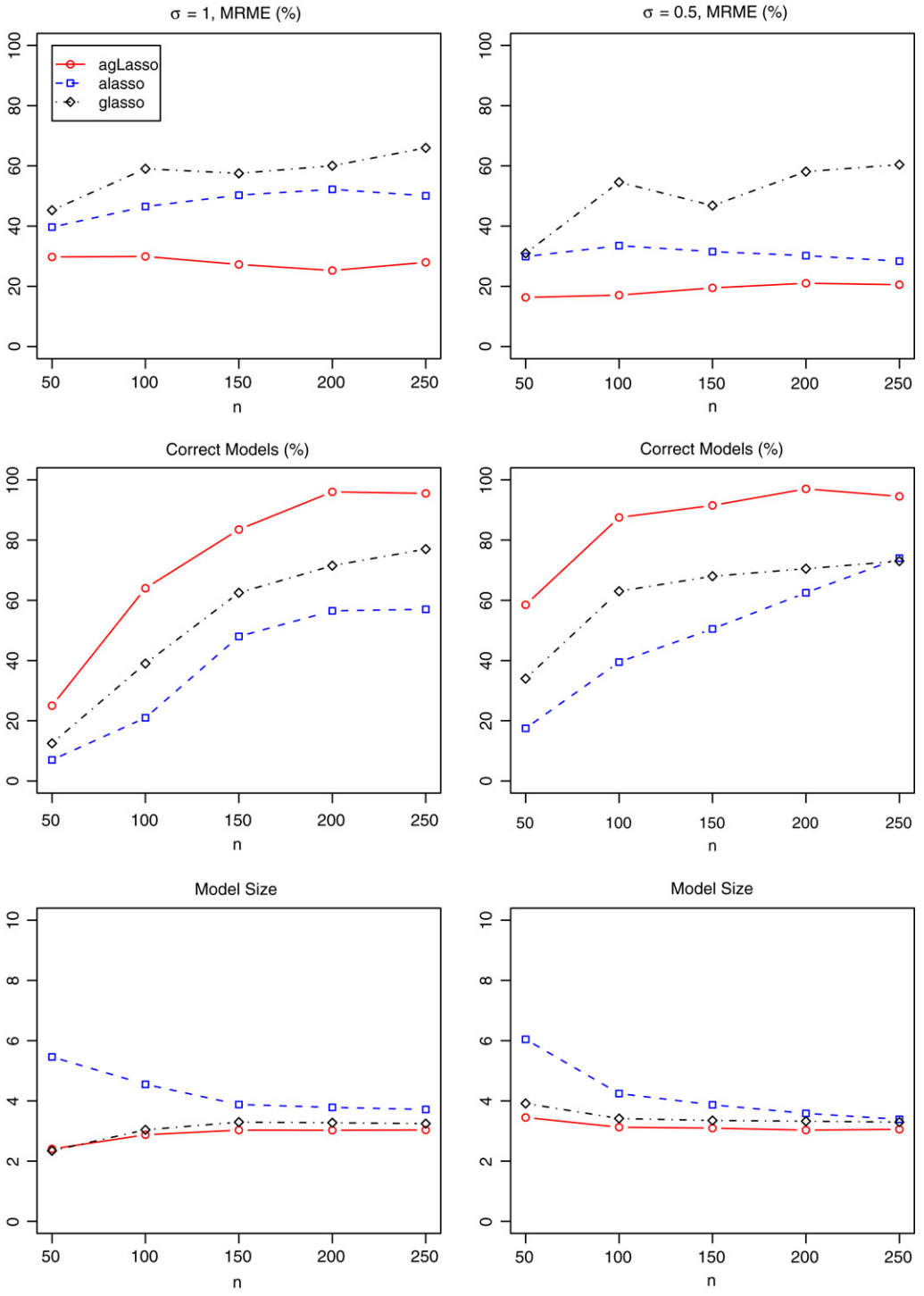

Fig. 2. Model 1 with BIC as the selection criterion.

Appendix

Proof of Theorem 1. Note that the agLasso objective function $Q(\beta)$ is a strictly convex function. Hence, as long as we can show that there is a local minimizer of (2.1), which is $\sqrt{n}$-consistent, then by the global convexity of (2.1), one knows immediately that such a local minimizer must be $\hat{\beta}$. Hence, the $\sqrt{n}$-consistency of $\hat{\beta}$ is established. Following a similar idea in Fan and Li (2001), the existence of a $\sqrt{n}$-consistent local minimizer is implied by the fact that for any $\epsilon > 0$, there is a

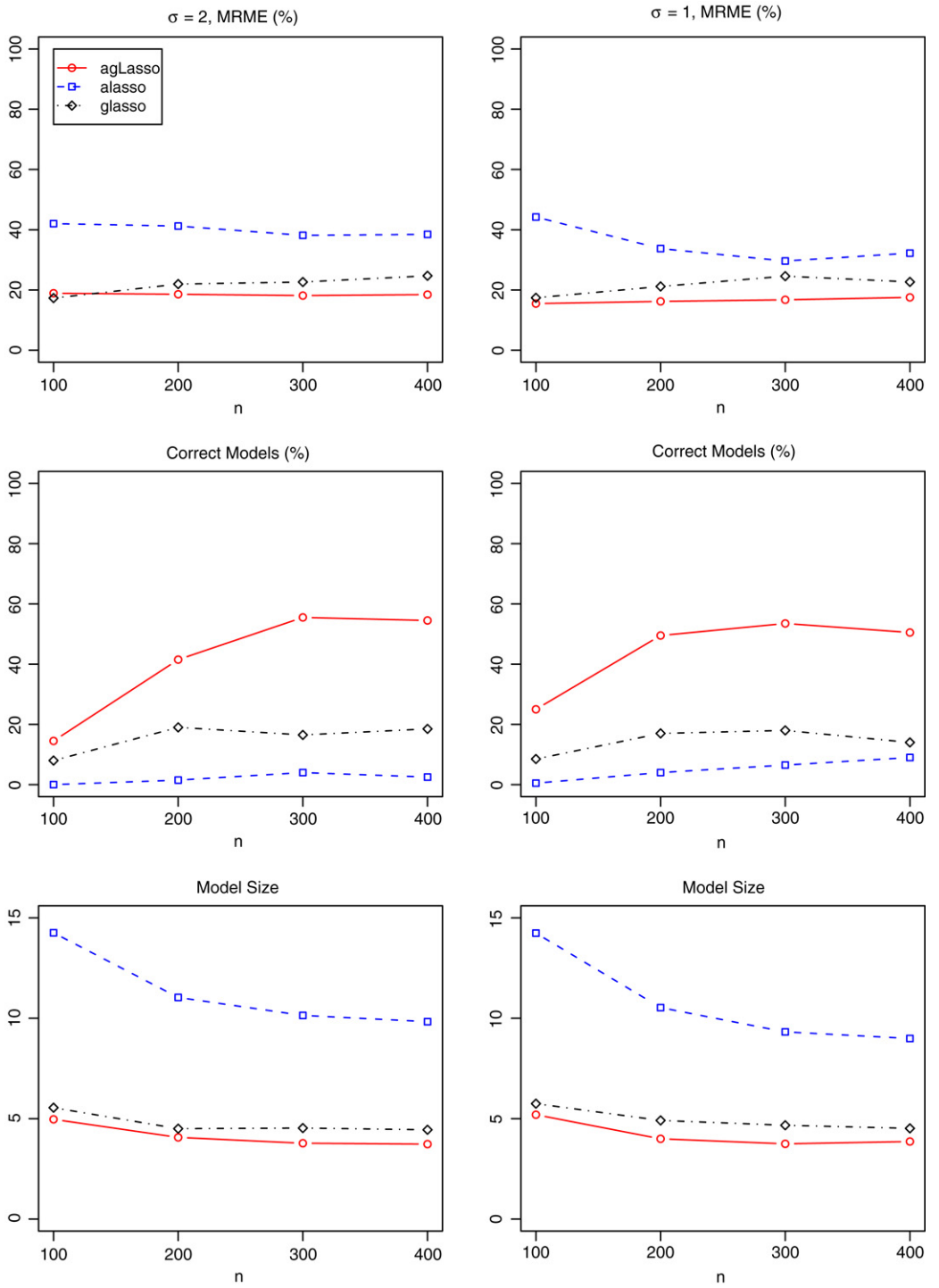

Fig. 3. Model 2 with Cp as the selection criterion.

sufficiently large constant C, such that

$$\liminf_n P\left\{\inf_{u\in\mathcal{R}^d:\|u\|=C} Q(\beta+n^{-1/2}u) > Q(\beta)\right\} > 1-\epsilon. \tag{A.1}$$

For simplicity, define the response vector as $Y=(y_1,\ldots,y_n)^\top$ and the design matrix as $X=(x_1,\ldots,x_n)^\top$. It follows then that

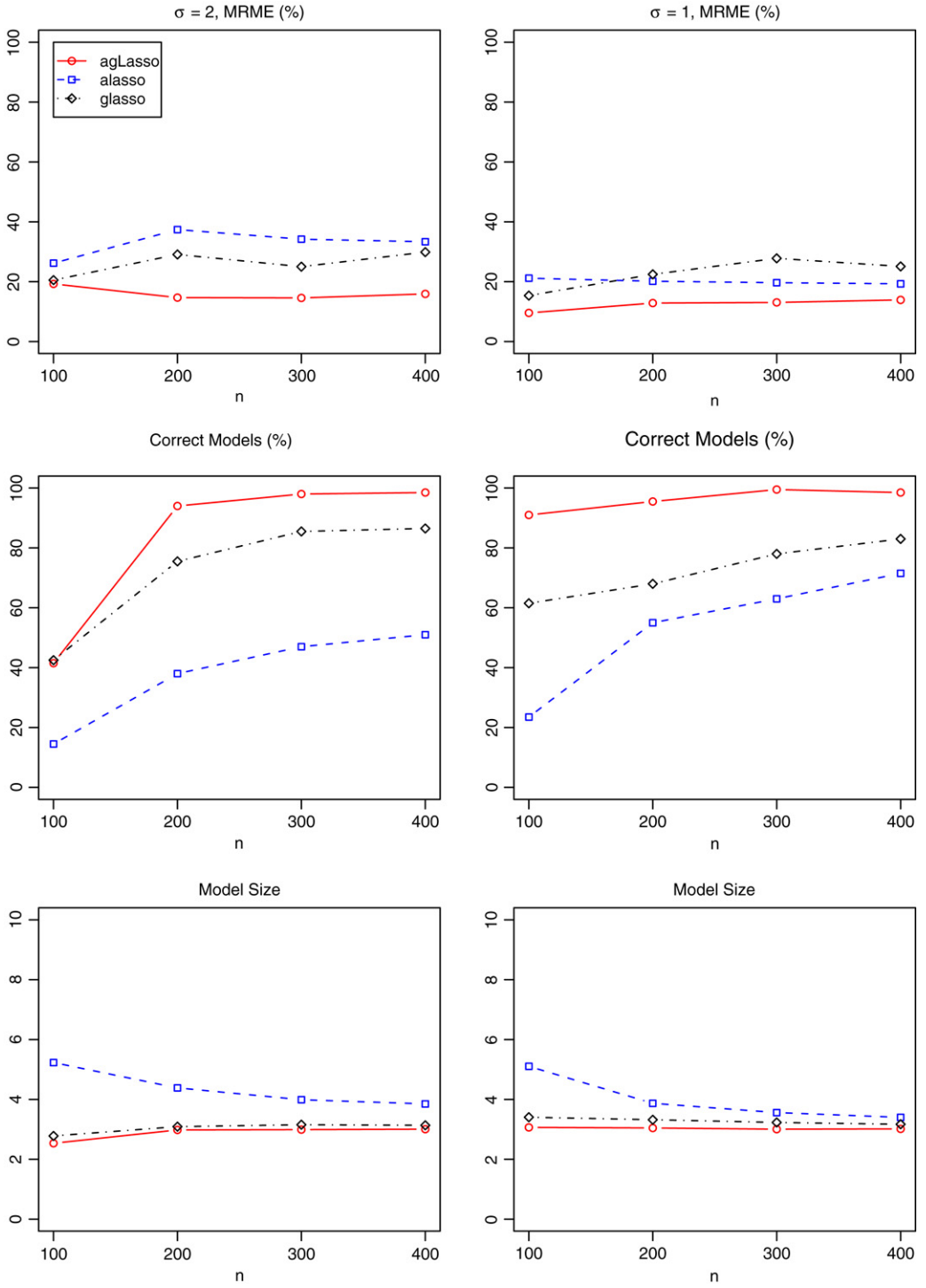

Fig. 4. Model 2 with BIC as the selection criterion.

$$Q(\beta + n^{-1/2}u) - Q(\beta) = \frac{1}{2}\|Y - X(\beta + n^{-1/2}u)\|^2 + n\sum_{j=1}^{p}\lambda_j\|\beta_j + n^{-1/2}u\| - \frac{1}{2}\|Y - X\beta\|^2 - n\sum_{j=1}^{p}\lambda_j\|\beta_j\|$$

$$= \frac{1}{2}u^\top\left(\frac{1}{n}X^\top X\right)u - u^\top\left(\frac{1}{\sqrt{n}}X^\top(Y - X\beta)\right) + n\sum_{j=1}^{p}\lambda_j\|\beta_j + n^{-1/2}u\| - n\sum_{j=1}^{p}\lambda_j\|\beta_j\|$$

$$= \frac{1}{2}u^\top\left(\frac{1}{n}X^\top X\right)u - u^\top\left(\frac{1}{\sqrt{n}}X^\top(Y - X\beta)\right) + n\sum_{j=1}^{p}\lambda_j\|\beta_j + n^{-1/2}u\| - n\sum_{j=1}^{p_0}\lambda_j\|\beta_j\| \quad \text{(A.2)}$$

$$\geq \frac{1}{2}u^\top\left(\frac{1}{n}X^\top X\right)u - u^\top\left(\frac{1}{\sqrt{n}}X^\top(Y - X\beta)\right) + n\sum_{j=1}^{p_0}\lambda_j(\|\beta_j + n^{-1/2}u\| - \|\beta_j\|)$$

$$\geq \frac{1}{2}u^\top\left(\frac{1}{n}X^\top X\right)u - u^\top\left(\frac{1}{\sqrt{n}}X^\top(Y - X\beta)\right) - p_0(\sqrt{n}a_n)\|u\|, \quad \text{(A.3)}$$

where equality (A.2) holds because $\beta_j = 0$ for any $j > p_0$ according to the model assumption. Furthermore, according to the theorem's condition, we know that $\sqrt{n}a_n = o_p(1)$, hence, the third term in (A.3) is $o_p(1)$. On the other hand, the first term converges in probability to $u^\top \mathrm{cov}(x)u$, which is a quadratic function in u. Lastly, the second term in (A.3) is linear in u with a $O_p(1)$ coefficient. Therefore, when C is sufficiently large, the first term dominates the other two terms with an arbitrarily large probability. This implies (A.1) and completes the proof. □

Proof of Theorem 2. Without loss of generality, we show in detail that $P(\hat{\beta}_p = 0) \to 1$. Then, the same argument can be used to show that $P(\hat{\beta}_j = 0) \to 1$ for any $p_0 < j < p$, which implies immediately that $P(\hat{\beta}_b = 0) \to 1$. For a better discussion, we define X_{-p} be a $n \times (d - d_p)$ matrix with the ith row given by $(x_{i1}^\top, \ldots, x_{i(p-1)}^\top)^\top$, the design matrix without the pth factor. Similarly, we define X_p to be the $n \times d_p$ design matrix with the ith row given by $x_{ip}^\top$. Next, we define $\beta_{-p} = (\beta_1^\top, \ldots, \beta_{p-1})^\top$ and let $\hat{\beta}_{-p}$ be its associated estimator. Note that if $\hat{\beta}_p \neq 0$, then the penalty function $\|\hat{\beta}_p\|$ becomes a differentiable function with respect to its components. Therefore, $\hat{\beta}_p$ must be the solution of the following normal equation

$$0 = \frac{1}{\sqrt{n}}X_p^\top\left(Y - X_{-p}\hat{\beta}_{-p} - X_p\hat{\beta}_p\right) + \sqrt{n}\lambda_p\frac{\hat{\beta}_p}{\|\hat{\beta}_p\|}$$

$$= \frac{1}{\sqrt{n}}X_p^\top(Y - X\beta) + \left(\frac{1}{n}X_p^\top X_{-p}\right)\sqrt{n}\left(\beta_{-p} - \hat{\beta}_{-p}\right) + \left(\frac{1}{n}X_p^\top X_p\right)\sqrt{n}\left(\beta_p - \hat{\beta}_p\right) + \sqrt{n}\lambda_p\frac{\hat{\beta}_p}{\|\hat{\beta}_p\|}, \quad \text{(A.4)}$$

where the first term in (A.4) is of the order $O_p(1)$, and the second and the third terms are also of the same order because $\beta_{-p} - \hat{\beta}_{-p} = O_p(n^{-1/2})$ and $\beta_p - \hat{\beta}_p = O_p(n^{-1/2})$ according to Theorem 1. Next note that if $\hat{\beta}_p \neq 0$, then there must exist a k such that $|\hat{\beta}_{pk}| = \max\{|\hat{\beta}_{pk'}| : 1 \leq k' \leq d_p\}$. Without loss of generality we can assume that $k = 1$, then we must have $|\hat{\beta}_{p1}|/\|\hat{\beta}_p\| \geq 1/\sqrt{d_p} > 0$. In addition to that, note that $\sqrt{n}\lambda_p \geq \sqrt{n}b_n \to \infty$. Therefore, we know that $\sqrt{n}\lambda_p\hat{\beta}_{pk}/\|\hat{\beta}_p\|$ dominates the first three terms in (A.4) with probability tending to one. This simply means that (A.4) cannot be true as long as the sample size is sufficiently large. As a result, we can conclude that with probability tending to one, the estimate $\hat{\beta}_p$ must be in a position where $\|\hat{\beta}_p\|$ is not differentiable. Hence, $\hat{\beta}_p$ has to be exactly 0. This completes the proof. □

Proof of Theorem 3. Based on the results of Theorems 1 and 2, we know that, with probability tending to one, we must have $\hat{\beta}_j \neq 0$ for $j \leq p_0$ and $\hat{\beta}_j = 0$ for $j > p_0$. Then, we know that, with probability tending to one, $\hat{\beta}_a$ must be the solution of the following normal equation

$$\frac{1}{n}X_a^\top(Y - X_a\hat{\beta}_a) + D(\hat{\beta}_a) = 0,$$

where $D(\hat{\beta}_a) = (\lambda_1\hat{\beta}_1^\top/\|\hat{\beta}_1\|, \ldots, \lambda_{p_0}\hat{\beta}_{p_0}^\top/\|\hat{\beta}_{p_0}\|)^\top$. It then follows that

$$\sqrt{n}(\hat{\beta}_a - \beta_a) = \left(\frac{1}{n}X_a^\top X_a\right)^{-1}\left(\frac{1}{\sqrt{n}}X_a^\top(Y - X_a\beta_a) + \sqrt{n}D(\hat{\beta}_a)\right). \quad \text{(A.5)}$$

Due to the fact that $\sqrt{n}\lambda_j \leq \sqrt{n}a_n \to_p 0$ for any $j \leq p_0$ and $|\hat{\beta}_{jk}|/\|\hat{\beta}_j\| < 1$ for any $1 \leq k \leq d_j$, we know that $D(\hat{\beta}_a) = o_p(n^{-1/2})$. Therefore, (A.5) can be further written as

$$\sqrt{n}(\hat{\beta}_a - \beta_a) = \left(\frac{1}{n}X_a^\top X_a\right)^{-1}\left(\frac{1}{\sqrt{n}}X_a^\top(Y - X_a\beta_a)\right) + o_p(1) \to_d N(0, \Sigma_a).$$

The theorem's conclusion follows and this completes the proof. □

References

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. The Annals of Statistics 32, 407–489.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96, 1348–1360.
Fu, W.J., 1998. Penalized regression: The bridge versus the LASSO. Journal of Computational and Graphical Statistics 7, 397–416.
Knight, K., Fu, W., 2000. Asymptotics for lasso-type estimators. The Annals of Statistics 28, 1356–1378.
Leng, C., Lin, Y., Wahba, G., 2006. A note on lasso and related procedures in model selection. Statistica Sinica 16, 1273–1284.
Tibshirani, R.J., 1996. Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society, Series B 58, 267–288.
Wang, H., Li, G., Jiang, G., 2007a. Robust regression shrinkage and consistent variable selection via the LAD-LASSO. Journal of Business and Economics Statistics 25, 347–355.
Wang, H., Li, G., Tsai, C.L., 2007b. Regression coefficient and autoregressive order shrinkage and selection via lasso. Journal of Royal Statistical Society, Series B 69, 63–78.
Wang, H., Li, R., Tsai, C.-L., 2007c. Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika 94, 553–568.
Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B 68, 49–67.
Yuan, M., Lin, Y., 2007. On the nonnegative garrote estimator. Journal of the Royal Statistical Society, Series B 69, 143–161.
Zhang, H.H., Lu, W., 2007. Adaptive lasso for Cox's proportional hazard model. Biometrika 94, 691–703.
Zou, H., 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101, 1418–1429.