

# A Note on Adaptive Group Lasso

Hansheng Wang and Chenlei Leng  
*Peking University & National University of Singapore*

July 17, 2006.

## 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.

KEY WORDS: Adaptive Group Lasso; Adaptive Lasso; Group Lasso; Lasso

## 1 Introduction

Since its first proposal by Tibshirani (1996), the *least absolute shrinkage and selection operator* (lasso) has generated much interest in 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 & 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 & 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. Then, the article concludes with a short discussion in Section 6. All technical details are presented in the Appendix.

## 2 Adaptive Group Lasso

### 2.1 Model and Notations

Let  $(x_1, y_1), \dots, (x_n, y_n)$  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  $x_i = (x_{i1}^\top, \dots, x_{ip}^\top)^\top$ , where  $x_{ij} = (x_{ij1}, \dots, x_{ijd_j})^\top \in \mathcal{R}^{d_j}$  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 order to model the dependence relationship between the response  $y_i$  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  $\beta_j = (\beta_{j1}, \dots, \beta_{jd_j})^\top \in \mathcal{R}^{d_j}$  is the regression coefficient vector associated with the  $j$ th factor and  $\beta$  is defined to be  $\beta = (\beta_1^\top, \dots, \beta_p^\top)^\top$ . Without loss of generality, we assume that only the first  $p_0 \leq p$  factors are relevant (i.e.,  $\|\beta_j\| \neq 0$  for  $j \leq p_0$  and  $\|\beta_j\| = 0$  for  $j > p_0$ ).

### 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

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

where  $\|\cdot\|$  stands for the typical  $L_2$  norm. 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  $\lambda$  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)

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

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  $\lambda_j$  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,  $\lambda_j$ ). 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

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

where  $\tilde{\beta} = (\tilde{\beta}_1^\top, \dots, \tilde{\beta}_p^\top)^\top$  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 originally  $p$ -dimensional tuning parameter selection problem for  $(\lambda_1, \dots, \lambda_p)$  reduces to a univariate problem for  $\lambda$  only. Thereafter, any appropriate selection method can be

used. In our numerical studies, the following selection criteria were considered:

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

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

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

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  $a_n = \max\{\lambda_j, j \leq p_0\}$  and  $b_n = \min\{\lambda_j, j > p_0\}$ .

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

Note that “ $\rightarrow_p$ ” 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  $\beta_a = (\beta_1^\top, \dots, \beta_{p_0}^\top)^\top$  to be the vector containing all the relevant factors, and let  $\beta_b = (\beta_{p_0+1}^\top, \dots, \beta_p^\top)^\top$  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  $\sqrt{n}(\tilde{\beta}_a - \beta_a) \rightarrow_d N(0, \Sigma_a)$  where  $\Sigma_a$  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  $\sqrt{na_n} \rightarrow_p 0$  and  $\sqrt{nb_n} \rightarrow_p \infty$ , then  $P(\hat{\beta}_b = 0) \rightarrow 1$ .*

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 non-zero coefficients. Both Theorem 1 and Theorem 2 imply that agLasso does have the ability to identify the true model consistently.

**Theorem 3.** (*Oracle Property*) If  $\sqrt{na_n} \rightarrow 0$  and  $\sqrt{nb_n} \rightarrow \infty$ , then  $\sqrt{n}(\hat{\beta}_a - \beta_a) \rightarrow_d N(0, \Sigma_a)$ .

Note that “ $\rightarrow_d$ ” denotes convergence in distribution. By Theorem 3, we know that, as long as the conditions  $\sqrt{na_n} \rightarrow 0$  and  $\sqrt{nb_n} \rightarrow \infty$  are satisfied, the resulting estimator is as efficient as *oracle*. On the other hand, one can verify easily that, as long as  $n^{1/2}\lambda \rightarrow 0$  and  $n^{(1+\gamma)/2}\lambda \rightarrow \infty$ , the theorem conditions  $\sqrt{na_n} \rightarrow 0$  and  $\sqrt{nb_n} \rightarrow \infty$  are satisfied.

## 4 Simulation Study

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 obtained from Yuan and Lin (2006).

*Example 1.* In this example, 15 latent variables  $Z_1, \dots, Z_{15}$  were generated according to a zero mean multivariate normal distribution, whose covariance between  $Z_i$  and  $Z_j$  was fixed to be  $0.5^{|i-j|}$ . Subsequently,  $Z_i$  was 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

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

where  $I(\cdot)$  is the indicator function and the residual  $\epsilon$  was normally distributed with mean 0 and standard deviation  $\sigma$ . 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) was summarized (Fan and Li, 2001). For each selection method, the percentage of the 200 simulated datasets, at which the true model is correctly identified, was computed. Lastly, the average model size (i.e., the number of factors) were 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 Figure 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.

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

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

where  $\epsilon \sim N(0, \sigma^2)$ . 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 Figure 3 and Figure 4. As one can see, the results are very similar to that of Example 1.

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	4.55	5.89	1.917	1.937	1.932
GCV	4.51	4.55	5.89	1.917	1.937	1.932
AIC	4.51	4.57	5.90	1.917	1.936	1.930
BIC	4.06	4.07	4.95	1.936	1.958	1.956

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

## 6 Discussion

In this article, we propose the agLasso method for adaptive grouped variable selection. Both our simulation and real data experience suggest that agLasso can outperform aLasso and gLasso

substantially. Our preliminary results are rather encouraging and the extension to generalized linear models should be straightforward.

## 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 that fact that for any  $\epsilon > 0$ , there is a sufficiently large constant  $C$ , such that

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

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

$$\begin{aligned} & Q(\beta + n^{-1/2}u) - Q(\beta) \\ &= \frac{1}{2} \|Y - X(\beta + n^{-1/2}u)\|^2 + n \sum_{j=1}^p \lambda_j \|\beta_j + n^{-1/2}u\| - \frac{1}{2} \|Y - X\beta\|^2 - n \sum_{j=1}^p \lambda_j \|\beta_j\| \\ &= \frac{1}{2} u^\top \left( \frac{1}{n} X^\top X \right) u - u^\top \left( \frac{1}{\sqrt{n}} X^\top (Y - X\beta) \right) + n \sum_{j=1}^p \lambda_j \|\beta_j + n^{-1/2}u\| - n \sum_{j=1}^p \lambda_j \|\beta_j\| \\ &= \frac{1}{2} u^\top \left( \frac{1}{n} X^\top X \right) u - u^\top \left( \frac{1}{\sqrt{n}} X^\top (Y - X\beta) \right) + n \sum_{j=1}^p \lambda_j \|\beta_j + n^{-1/2}u\| - n \sum_{j=1}^{p_0} \lambda_j \|\beta_j\| \quad (\text{A.2}) \\ &\geq \frac{1}{2} u^\top \left( \frac{1}{n} X^\top X \right) u - u^\top \left( \frac{1}{\sqrt{n}} X^\top (Y - X\beta) \right) + n \sum_{j=1}^{p_0} \lambda_j (\|\beta_j + n^{-1/2}u\| - \|\beta_j\|) \\ &\geq \frac{1}{2} u^\top \left( \frac{1}{n} X^\top X \right) u - u^\top \left( \frac{1}{\sqrt{n}} X^\top (Y - X\beta) \right) - p_0(\sqrt{n}a_n)\|u\|, \quad (\text{A.3}) \end{aligned}$$

where the 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  $u^\top \text{cov}(x)u$ , which is a quadratic function in  $u$ . Last, 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 arbitrary large probability. This implies (A.1) and completes the proof.

**Proof of Theorem 2:** Without loss of generality, we show in detail that  $P(\hat{\beta}_p = 0) \rightarrow 1$ .

Then, the same argument can be used to show that  $P(\hat{\beta}_j = 0) \rightarrow 1$  for any  $p_0 < j < p$ , which implies immediately that  $P(\hat{\beta}_b = 0) \rightarrow 1$ . For a better discussion, we define  $X_{-p}$  be a  $n \times (d - d_p)$  matrix with the  $i$ th row given by  $(x_{i1}^\top, \dots, x_{i(p-1)}^\top)^\top$ , the design matrix without the  $p$ th factor. Similarly, we define  $X_p$  to be the  $n \times d_p$  design matrix with the  $i$ th row given by  $x_{ip}^\top$ . Next, we define  $\beta_{-p} = (\beta_1^\top, \dots, \beta_{p-1}^\top)^\top$  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

$$\begin{aligned} 0 &= \frac{1}{\sqrt{n}} X_p^\top (Y - X_{-p} \hat{\beta}_{-p} - X_p \hat{\beta}_p) + \sqrt{n} \lambda_p \frac{\hat{\beta}_p}{\|\hat{\beta}_p\|} \\ &= \frac{1}{\sqrt{n}} X_p^\top (Y - X \beta) + \left( \frac{1}{n} X_p^\top X_{-p} \right) \sqrt{n} (\beta_{-p} - \hat{\beta}_{-p}) \\ &\quad + \left( \frac{1}{n} X_p^\top X_p \right) \sqrt{n} (\beta_p - \hat{\beta}_p) + \sqrt{n} \lambda_p \frac{\hat{\beta}_p}{\|\hat{\beta}_p\|}, \end{aligned} \quad (\text{A.4})$$

where the first term in (A.4) is of the order  $O_p(1)$ , and the second and the third term 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  $\sqrt{n} \lambda_p \geq \sqrt{n} b_n \rightarrow \infty$ . 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 Theorem 1 and Theorem 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  $D(\hat{\beta}_a) = (\lambda_1 \hat{\beta}_1^\top / \|\hat{\beta}_1\|, \dots, \lambda_{p_0} \hat{\beta}_{p_0}^\top / \|\hat{\beta}_{p_0}\|)^\top$ . It follows then

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

Due to the fact that  $\sqrt{n} \lambda_j \leq \sqrt{n} a_n \rightarrow_p 0$  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

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



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

## References

- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least angle regression," *The Annals of Statistics*, 32, 407–489.
- 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.
- Fu, W. J. (1998), "Penalized regression: the bridge versus the LASSO," *Journal of Computational and Graphical Statistics*, 7, 397–416.
- Knight, K. and Fu, W. (2000), "Asymptotics for lasso-type estimators," *The Annals of Statistics*, 28, 1356–1378.
- Leng, C., Lin, Y., and 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., and Jiang, G. (2007a), "Robust regression shrinkage and consistent variable selection via the LAD-LASSO," *Journal of Business and Economics Statistics*, To Appear.
- Wang, H., Li, G., and Tsai, C. L. (2007b), "Regression coefficient and autoregressive order shrinkage and selection via lasso," *Journal of Royal Statistical Society, Series B*, To Appear.
- Yuan, M. and Lin, Y. (2006), "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society, Series B*, 68, 49–67.
- (2007), "On the nonnegative garrote estimator," *Journal of the Royal Statistical Society, Series B*, To appear.
- Zhang, H. H. and Lu, W. (2007), "Adaptive lasso for Cox's proportional hazard model," *Biometrika*, To appear.
- Zou, H. (2006), "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Figure 1: Model 1 with Cp as the Selection Criterion

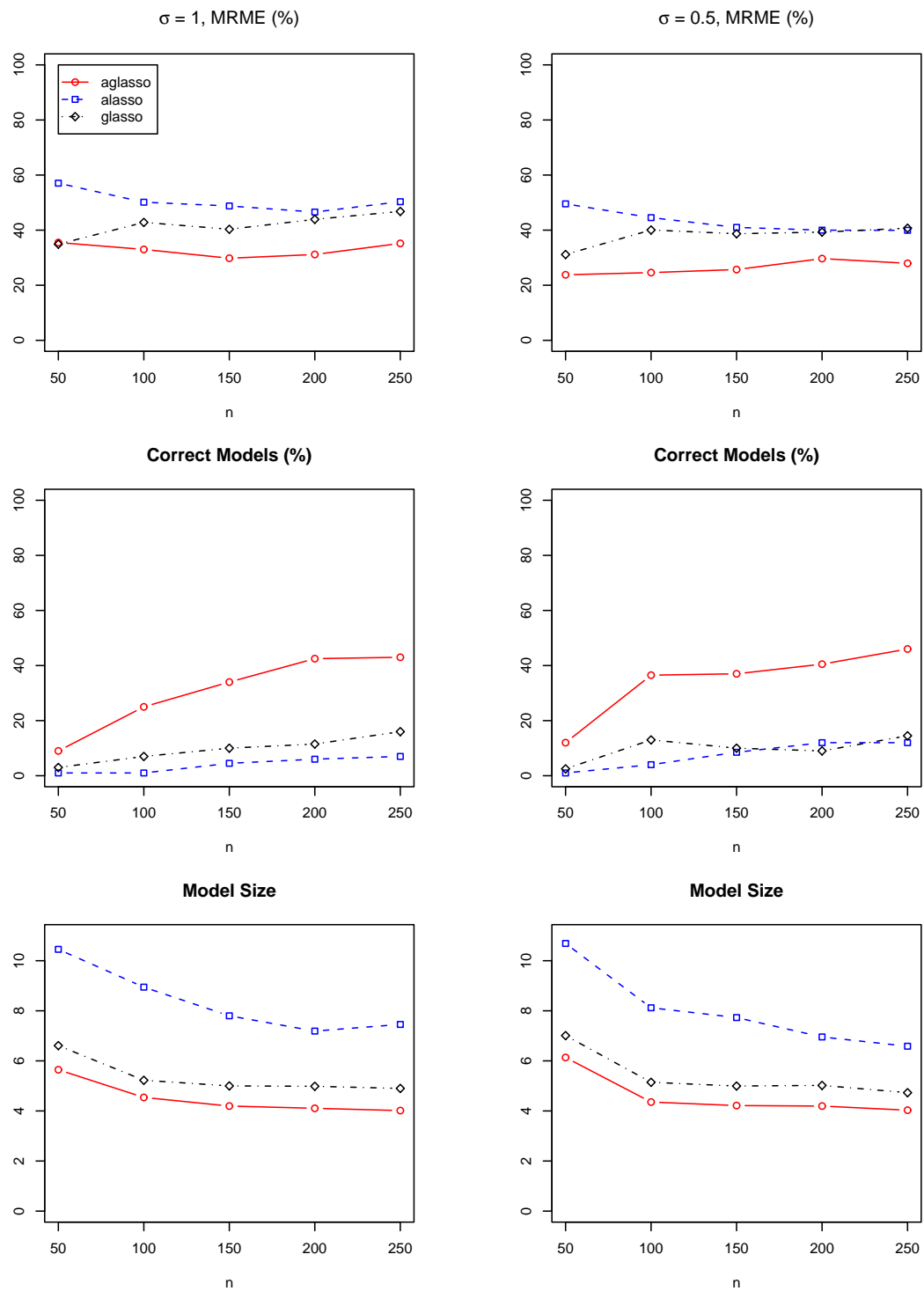


Figure 2: Model 1 with BIC as the Selection Criterion

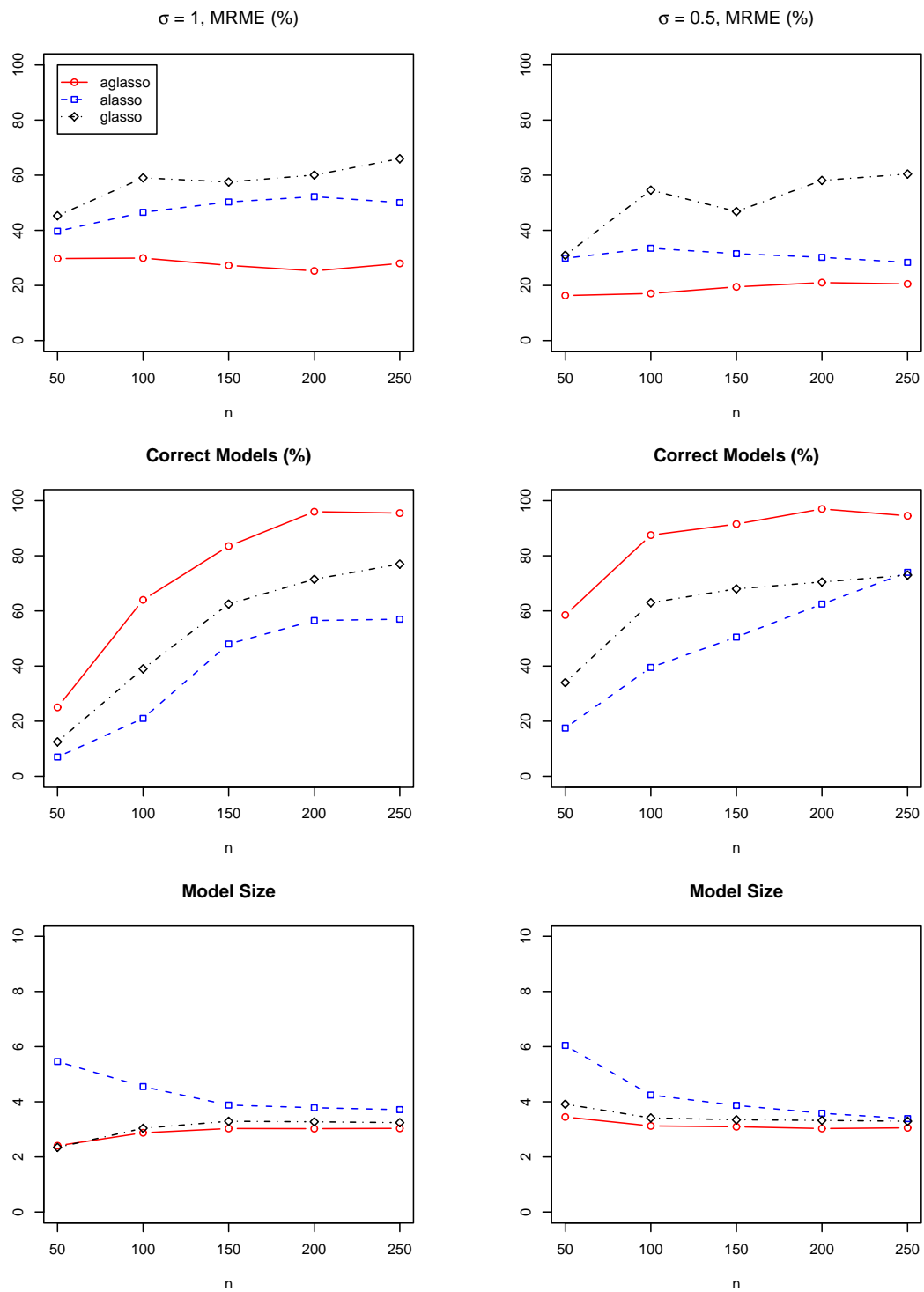


Figure 3: Model 2 with Cp as the Selection Criterion

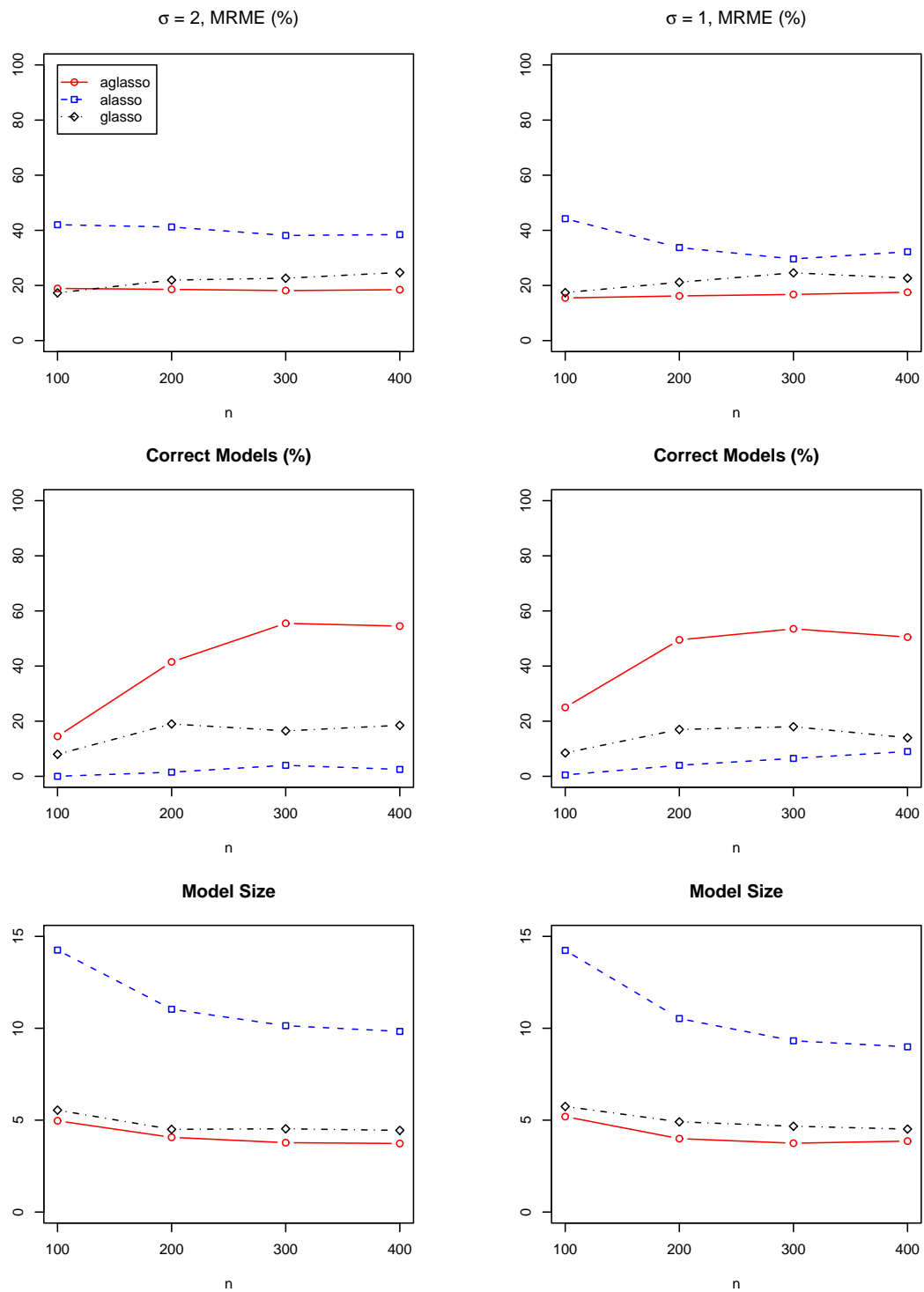


Figure 4: Model 2 with BIC as the Selection Criterion

