

Unified LASSO Estimation via Least Squares Approximation

Hansheng Wang and Chenlei Leng

Peking University & National University of Singapore

First version: May 25, 2006.

Revised on March 23, 2007.

Abstract

We propose a method of least squares approximation (LSA) for unified yet simple LASSO estimation. Our general theoretical framework includes ordinary least squares, generalized linear models, quantile regression, and many others as special cases. Specifically, LSA can transfer many different types of LASSO objective functions into their asymptotically equivalent least-squares problems. Thereafter, the standard asymptotic theory can be established and the LARS algorithm can be applied. In particular, if the adaptive LASSO penalty and a BIC-type tuning parameter selector are used, the resulting LSA estimator can be as efficient as oracle. Extensive numerical studies confirm our theory.

KEY WORDS: Adaptive LASSO; BIC; LASSO; Least Angle Regression; Least Squares Approximation; Microarray Data; Oracle Property; Solution Path

Hansheng Wang is Associate Professor at Guanghua School of Management, Peking University, Beijing, P. R. China, 100871 (hansheng@gsm.pku.edu.cn). Chenlei Leng is Assistant Professor at Department of Statistics and Applied Probability, National University of Singapore, Singapore, (stalc@nus.edu.sg). The research of Leng is supported partially by NUS grant R-155-050-053-133.

Acknowledgement: The authors are very grateful to the Editor, the Associate Editor, and two referees for their careful reading and constructive comments. The authors are also very grateful to the Associate Editor, Prof. Bruce Brown, Prof. Chih-Ling Tsai, and Prof. Alec Stephenson for many detailed editorial suggestions, which led to a better presentation of the paper.

1. INTRODUCTION

The least absolute shrinkage and selection operator (LASSO; Tibshiranni, 1996) is a useful and well-studied approach to the problem of variable selection (Knight and Fu, 2000; Fan and Li, 2001; Leng et al., 2006; Wang et al., 2007; Yuan and Lin, 2007). Compared with traditional estimation methods, LASSO's major advantage is its simultaneous execution of both parameter estimation and variable selection (Tibshirani, 1996; Fan and Li, 2001). In particular, allowing an adaptive amount of shrinkage for each regression coefficient results in an estimator as efficient as oracle (Zou, 2006; Wang et al., 2007). Furthermore, LASSO has very good computational properties. In the ordinary least squares (OLS) setting, Osborne et al. (2000) and Efron et al. (2004) showed that the solution path of LASSO is piecewise linear and moves in a very predictable manner. Taking advantage of this property, Efron et al. (2004) developed least angle regression (LARS), which finds the entire solution path of LASSO at the same computational cost as one single OLS fit. Similar path-finding algorithms were also developed for other optimization problems by Rosset (2004), Zhao and Yu (2004), and Park and Hastie (2006a).

These path-finding algorithms are computationally fast and have been proved very useful in both theory and practice. Furthermore, with reasonable programming effort, these algorithms can be modified easily to obtain solution paths for many other objective functions. In many real applications, however, the field practitioners may have considerable experience as software users but not as statistical programmers. For those researchers, the programming effort demanded by modifying any existing path-finding algorithm could be challenging. We believe that such a difficulty may limit more extensive application of the novel LASSO and LARS methodologies. Hence, there is a need for unified and simple treatment of various LASSO methods.

LASSO's theoretical properties have been well studied for OLS, generalized linear models (GLMs), and Cox models. But rather less is known about LASSO for many other regression models, e.g., quantile regression (Koenker and Bassett, 1978). Theoretically, similar asymptotic and numerical theories can be established for these models in a case-by-case manner. Such a task is not only tedious but also quite demanding. Consequently, an interesting question arises: is it possible to include many different regression models into one unified theoretical framework?

As a simple solution, we propose a method of least squares approximation (LSA) for unified LASSO estimation. Our general theoretical framework includes ordinary least squares, generalized linear models, quantile regression, and many others as special cases. Specifically, LSA is able to transfer many different types of LASSO objective functions into their asymptotically equivalent least-squares problems. Thereafter, the standard asymptotic theory can be established and the LARS algorithm can be applied. In particular, if the adaptive LASSO penalty (Zou, 2006; Wang et al., 2007) and a BIC-type tuning parameter selector (Wang et al., 2007) are used, the resulting LSA estimator can be as efficient as oracle.

The rest of the article is organized as follows. Section 2 introduces the LSA method, whose main theoretical properties are discussed in Section 3. Extensive numerical studies are reported in Section 4 and the article concludes with some discussion in Section 5. The Appendix contains the technical details.

2. THE LEAST SQUARES APPROXIMATION

2.1. The Model and Notations

Let $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ be n independent and identically distributed random vectors, where $y_i \in \mathbb{R}^1$ is the response of interest and $\mathbf{x}_i = (x_{i1}, \dots, x_{id})^\top \in \mathbb{R}^d$ is the

d -dimensional predictor. Assume that $\boldsymbol{\beta} = (\beta_1, \dots, \beta_d)^\top \in \mathbb{R}^d$ is a parameter of interest and $\mathcal{L}_n(\boldsymbol{\beta})$ is a plausible loss function, whose global minimizer $\tilde{\boldsymbol{\beta}} = \operatorname{argmin} \mathcal{L}_n(\boldsymbol{\beta})$ is a natural estimate of $\boldsymbol{\beta}$. For example, if $\mathcal{L}_n(\boldsymbol{\beta})$ is the least squares function, $\tilde{\boldsymbol{\beta}}$ is the usual OLS estimate. If $\mathcal{L}_n(\boldsymbol{\beta})$ is the negative log-likelihood function, $\tilde{\boldsymbol{\beta}}$ is the maximum likelihood estimator (MLE). Assume further that $\tilde{\boldsymbol{\beta}}$ is \sqrt{n} -consistent and asymptotically normal, which is the case for most common linear regression methods. Thus, $\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \rightarrow_d N(0, \boldsymbol{\Sigma})$ for some true $\boldsymbol{\beta}_0 = (\beta_{01}, \dots, \beta_{0d})^\top \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is referred to as the asymptotic covariance matrix of $\tilde{\boldsymbol{\beta}}$.

For simultaneous parameter estimation and variable selection, we consider the following adaptive LASSO objective function (Zou, 2006; Wang et al., 2007)

$$n^{-1} \mathcal{L}_n(\boldsymbol{\beta}) + \sum_{j=1}^d \lambda_j |\beta_j|. \quad (2.1)$$

The adaptive LASSO penalty utilized in (2.1) is slightly different from the traditional one (Tibshirani, 1996), in which the same regularization parameter is used for every regression coefficient. As noted by a number of authors, because the traditional LASSO uses the same amount of shrinkage for each regression coefficient, the resulting estimator cannot be as efficient as oracle (Fan and Li, 2001) and the selection results can be inconsistent (Leng et al., 2006; Zou, 2006; Yuan and Lin, 2007). As a simple solution, both Zou (2006) and Wang et al. (2007) proposed the above adaptive LASSO penalty, whereby different amounts of shrinkage are used for different regression coefficients. Intuitively, if larger amounts of shrinkage are applied to the zero coefficients while smaller amounts are used for the nonzero ones, an estimator with improved efficiency can be obtained. This is referred to as the adaptive LASSO (Zou, 2006).

2.2. The Least Squares Approximation

To motivate the least squares approximation (LSA) method, we further assume that $\mathcal{L}_n(\boldsymbol{\beta})$ has continuous second order derivative with respect to $\boldsymbol{\beta}$. This assumption is used only to motivate the idea. Subsequently, we will see that LSA's final implementation is completely independent of this assumption. Recall that $\tilde{\boldsymbol{\beta}} \in \mathbb{R}^d$ is the unpenalized estimator obtained by minimizing $\mathcal{L}_n(\boldsymbol{\beta})$. A standard Taylor series expansion at $\tilde{\boldsymbol{\beta}}$ gives

$$n^{-1}\mathcal{L}_n(\boldsymbol{\beta}) \approx n^{-1}\mathcal{L}_n(\tilde{\boldsymbol{\beta}}) + n^{-1}\dot{\mathcal{L}}_n(\tilde{\boldsymbol{\beta}})^\top(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) + \frac{1}{2}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^\top \left\{ \frac{1}{n}\ddot{\mathcal{L}}_n(\tilde{\boldsymbol{\beta}}) \right\} (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}), \quad (2.2)$$

where $\dot{\mathcal{L}}_n(\cdot)$ and $\ddot{\mathcal{L}}_n(\cdot)$ are the first and second order derivatives of the loss function $\mathcal{L}_n(\cdot)$, respectively. Because $\tilde{\boldsymbol{\beta}}$ is the minimizer of $\mathcal{L}_n(\cdot)$, we know that $\dot{\mathcal{L}}_n(\tilde{\boldsymbol{\beta}}) = 0$. Hence, the approximation (2.2) can be simplified to

$$n^{-1}\mathcal{L}_n(\boldsymbol{\beta}) \approx n^{-1}\mathcal{L}_n(\tilde{\boldsymbol{\beta}}) + \frac{1}{2}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^\top \left\{ \frac{1}{n}\ddot{\mathcal{L}}_n(\tilde{\boldsymbol{\beta}}) \right\} (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}). \quad (2.3)$$

This implies that, locally around $\tilde{\boldsymbol{\beta}}$, the original loss function can be well approximated by the above least squares type objective function.

By ignoring the constant $\mathcal{L}_n(\tilde{\boldsymbol{\beta}})$ and the coefficient 1/2, the objective function (2.3) can be further simplified to $(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^\top \{n^{-1}\ddot{\mathcal{L}}_n(\tilde{\boldsymbol{\beta}})\}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})$. Very often the quantity $n^{-1}\ddot{\mathcal{L}}_n(\tilde{\boldsymbol{\beta}})$ is closely related to the asymptotic covariance of $\tilde{\boldsymbol{\beta}}$ (i.e., $\boldsymbol{\Sigma}$). Specifically, it is plausible that $E\{n^{-1}\ddot{\mathcal{L}}_n(\boldsymbol{\beta})\} \approx \boldsymbol{\Sigma}^{-1}$, and $\hat{\boldsymbol{\Sigma}}^{-1} = n^{-1}\ddot{\mathcal{L}}_n(\tilde{\boldsymbol{\beta}})$ is a natural estimate for $\boldsymbol{\Sigma}^{-1}$. This motivates us to consider the following least squares function

$$(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) \quad (2.4)$$

as a simple approximation to the original loss $n^{-1}\mathcal{L}_n(\boldsymbol{\beta})$. We refer to (2.4) as the least squares approximation (LSA). In some situations (e.g., quantile regression), the loss function $\mathcal{L}_n(\boldsymbol{\beta})$ is not sufficiently smooth. This condition usually does not prevent the unpenalized estimator $\tilde{\boldsymbol{\beta}}$ from being \sqrt{n} -consistent or asymptotically normal, nor does it prevent the existence of a consistent asymptotic covariance estimator $\hat{\boldsymbol{\Sigma}}$. Thus, even if $\mathcal{L}_n(\boldsymbol{\beta})$ is not sufficiently smooth, the LSA method defined in Section 2.3 below is still useful as long as $\hat{\boldsymbol{\Sigma}}$ is available.

2.3. The LSA Estimator

With LSA defined in (2.4), the original LASSO problem (2.1) can be rewritten as the following asymptotically equivalent least squares problem

$$Q(\boldsymbol{\beta}) = (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^\top \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) + \sum_{j=1}^d \lambda_j |\beta_j|, \quad (2.5)$$

whose global minimizer is given by $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} = (\hat{\beta}_{\lambda,1}, \dots, \hat{\beta}_{\lambda,d})^\top$ for $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_d)^\top$. In general, $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ is different from the adaptive LASSO estimator, which is obtained by minimizing (2.1). Hereafter, we refer to $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ as the LSA estimator and the minimizer of (2.1) (denoted by $\tilde{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$) as the adaptive LASSO (aLASSO) estimator.

The approximated LASSO objective function (2.5) is nothing but an L_1 -penalized least squares problem. Although the original motivation assumed certain smoothness conditions for $\mathcal{L}_n(\boldsymbol{\beta})$ (Taylor series expansion), the final implementation (2.5) is completely independent of this requirement. The final form requires only the existence of a consistent covariance matrix estimate $\hat{\boldsymbol{\Sigma}}$, which is the case for most existing regression models. In many situations, $\hat{\boldsymbol{\Sigma}}$ is even a standard output of many commonly used statistical packages. For example, both $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}$ are standard outputs from the R functions *lm* (OLS), *glm* (GLM), *coxph* (Cox's Model), etc. Together with the *lars* package

in \mathbb{R} , the proposed LSA method can be implemented easily. This further extends the practical applicability of the LASSO method and the LARS algorithm. Finally, because $\tilde{\boldsymbol{\beta}}$ utilized by LSA is not asymptotically efficient compared with oracle, one may expect that the resulting LSA estimator is not efficient either. However, as we shall see in the next section, the LSA estimator is indeed as efficient as oracle asymptotically, as long as the tuning parameters are selected appropriately.

3. THEORETICAL PROPERTIES

3.1. The Covariance Assumption

To study the properties of LSA, we first define some notation. Without loss of generality, assume that there exists a finite dimensional true model, for which only the first d_0 ($0 \leq d_0 \leq d$) predictors are relevant, i.e., $\boldsymbol{\beta}_{0j} \neq 0$ for $j \leq d_0$ and $\boldsymbol{\beta}_{0j} = 0$ for $j > d_0$. Let $\mathcal{S} = \{j_1, \dots, j_{d^*}\}$ denote an arbitrary candidate model, which contains the j_1 th, \dots , j_{d^*} th ($1 \leq d^* \leq d$) predictors, and let $\mathcal{S}_F = \{1, \dots, d\}$ and $\mathcal{S}_T = \{1, \dots, d_0\}$ denote the full and true model, respectively. An arbitrary candidate model \mathcal{S} is an underfitted model if $\mathcal{S} \not\supset \mathcal{S}_T$, or an overfitted model if $\mathcal{S} \supset \mathcal{S}_T$ and $\mathcal{S} \neq \mathcal{S}_T$. Let $|\mathcal{S}|$ denote the size of the model \mathcal{S} (i.e., the number of variables contained in \mathcal{S}). Lastly, denote $\tilde{\boldsymbol{\beta}}_{\mathcal{S}} \in \mathbb{R}^d$ as the corresponding unpenalized estimator, i.e.,

$$\tilde{\boldsymbol{\beta}}_{\mathcal{S}} = \operatorname{argmin}_{\{\boldsymbol{\beta} \in \mathbb{R}^d: \boldsymbol{\beta}_j = 0, \forall j \notin \mathcal{S}\}} \mathcal{L}_n(\boldsymbol{\beta}).$$

This implies that $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}_{\mathcal{S}_F}$. On the other hand, by minimizing the objective function (2.4), another estimator

$$\hat{\boldsymbol{\beta}}_{\mathcal{S}} = \operatorname{argmin}_{\{\boldsymbol{\beta} \in \mathbb{R}^d: \boldsymbol{\beta}_j = 0, \forall j \notin \mathcal{S}\}} (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^{\top} \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) \quad (3.1)$$

can be obtained. In general, $\tilde{\beta}_{\mathcal{S}} \neq \hat{\beta}_{\mathcal{S}}$. The estimator $\tilde{\beta}_{\mathcal{S}}$ is obtained by minimizing the loss function $\mathcal{L}_n(\cdot)$, while $\hat{\beta}_{\mathcal{S}}$ is produced by minimizing the LSA objective function (2.4). Although they are closely related, they are unlikely to be exactly the same. A similar difference also exists between $\tilde{\beta}_{\lambda}$ and $\hat{\beta}_{\lambda}$.

For an arbitrary vector $\alpha \in \mathbb{R}^d$, let $\alpha^{(\mathcal{S})} \in \mathbb{R}^{|\mathcal{S}|}$ denote the subvector of α associated with the candidate model \mathcal{S} . Thus, we have $\tilde{\beta} = \tilde{\beta}^{(\mathcal{S}_F)}$, $\beta_0^{(\mathcal{S}_T)} \neq 0$, and the oracle estimator $\tilde{\beta}_{\mathcal{S}_T}^{(\mathcal{S}_T)}$. In addition, assume that for any $\mathcal{S} \supset \mathcal{S}_T$,

$$\sqrt{n}(\tilde{\beta}_{\mathcal{S}}^{(\mathcal{S})} - \beta_0^{(\mathcal{S})}) \rightarrow_d N(0, \Sigma_{\mathcal{S}}) = N(0, \Omega_{\mathcal{S}}^{-1}) \quad (3.2)$$

for some positive definite matrix $\Sigma_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$, whose inverse is given by $\Omega_{\mathcal{S}} = \Sigma_{\mathcal{S}}^{-1}$. The notation implies that $\Sigma_{\mathcal{S}_F} = \Sigma$. Furthermore, the asymptotic covariance matrix of the oracle estimator $\tilde{\beta}_{\mathcal{S}_T}^{(\mathcal{S}_T)}$ is given by $\Sigma_{\mathcal{S}_T}$, i.e.,

$$\sqrt{n}(\tilde{\beta}_{\mathcal{S}_T}^{(\mathcal{S}_T)} - \beta_0^{(\mathcal{S}_T)}) \rightarrow_d N(0, \Sigma_{\mathcal{S}_T}) = N(0, \Omega_{\mathcal{S}_T}^{-1}). \quad (3.3)$$

Lastly, for an arbitrary $d \times d$ matrix \mathbf{A} , let $\mathbf{A}^{(\mathcal{S})}$ denote the submatrix associated with \mathcal{S} . We now state formally the following important covariance assumption, which is the key to establishing the oracle property.

THE COVARIANCE ASSUMPTION: For any $\mathcal{S} \supset \mathcal{S}_T$, we have $\Omega_{\mathcal{S}} = \Omega_{\mathcal{S}_F}^{(\mathcal{S})}$.

Simply speaking, the covariance assumption indicates that there exists a close relationship between the inverse asymptotic covariance matrix of an overfitted model (i.e., $\Omega_{\mathcal{S}}$) and that of the full model (i.e., $\Omega_{\mathcal{S}_F}$). This assumption is easily satisfied in many common regression problems.

Case 1. Consider the simplest situation, OLS regression with independent er-

rors: $\boldsymbol{\Omega}_{\mathcal{S}} = \sigma^{-2}\text{cov}(\mathbf{x}^{(\mathcal{S})})$ and σ^2 is the common variance of the error terms. Because $\text{cov}(\mathbf{x}^{(\mathcal{S})})$ is a submatrix of $\text{cov}(\mathbf{x}^{(\mathcal{S}_F)})$, the covariance assumption $\boldsymbol{\Omega}_{\mathcal{S}} = \boldsymbol{\Omega}_{\mathcal{S}_F}^{(\mathcal{S})}$ is satisfied. In fact, as long as the inverse asymptotic covariance matrix satisfies $\boldsymbol{\Omega}_{\mathcal{S}} = \tau\text{cov}(\mathbf{x}^{(\mathcal{S})})$ for some $\tau > 0$, the covariance assumption is satisfied.

Case 2. Consider a logistic regression model, where y is a binary response with the following conditional probability function

$$P(y = 1|x) = p(\mathbf{x}^{(\mathcal{S}_T)}) = \frac{\exp(\boldsymbol{\beta}_0^{(\mathcal{S}_T)\top} \mathbf{x}^{(\mathcal{S}_T)})}{1 + \exp(\boldsymbol{\beta}_0^{(\mathcal{S}_T)\top} \mathbf{x}^{(\mathcal{S}_T)})}.$$

For any overfitted model, the inverse asymptotic covariance matrix of the corresponding MLE (i.e., $\tilde{\boldsymbol{\beta}}_{\mathcal{S}}^{(\mathcal{S})}$) is $\boldsymbol{\Omega}_{\mathcal{S}} = \text{E}\{p(\mathbf{x}^{(\mathcal{S}_T)})(1 - p(\mathbf{x}^{(\mathcal{S}_T)}))\mathbf{x}^{(\mathcal{S})}\mathbf{x}^{(\mathcal{S})\top}\}$, a submatrix of $\boldsymbol{\Omega}_{\mathcal{S}_F} = \text{E}\{p(\mathbf{x}^{(\mathcal{S}_T)})(1 - p(\mathbf{x}^{(\mathcal{S}_T)}))\mathbf{x}^{(\mathcal{S}_F)}\mathbf{x}^{(\mathcal{S}_F)\top}\}$. Hence, the covariance assumption is satisfied. Similarly, for any regression model whose likelihood function is of the form $f(y, \mathbf{x}^\top \boldsymbol{\beta})$, the covariance assumption is satisfied.

Case 3. As the final example, consider least absolute deviation (LAD) regression, for which the loss function is not smooth. Assume that the response y depends on the predictor \mathbf{x} via the linear regression model $y = \mathbf{x}^\top \boldsymbol{\beta} + e$, where e is some random noise with median 0. For any model $\mathcal{S} \supset \mathcal{S}_T$, the LAD estimator $\hat{\boldsymbol{\beta}}_{\mathcal{S}}$ is \sqrt{n} -consistent and asymptotically normal. Its inverse asymptotic covariance matrix is given by $\boldsymbol{\Omega}_{\mathcal{S}} = 4f_e^2(0)\text{cov}(\mathbf{x}^{(\mathcal{S})})$, where $f_e(0)$ is the probability density of the random noise e at the origin (Pollard, 1991). Similar to Case 1, because $\text{cov}(\mathbf{x}^{(\mathcal{S})})$ is the submatrix of $\text{cov}(\mathbf{x}) = \text{cov}(\mathbf{x}^{(\mathcal{S}_F)})$ according to \mathcal{S} , it follows that $\boldsymbol{\Omega}_{\mathcal{S}} = \boldsymbol{\Omega}_{\mathcal{S}_F}^{(\mathcal{S})}$. Once again, the covariance assumption is satisfied. In fact, one can check that such a conclusion is also valid for a general quantile regression (Koenker and Bassett, 1978).

There certainly exist situations where the covariance assumption is violated. For

example, in Case 1, if the random noise depends on the predictor \mathbf{x} , the covariance assumption may be violated. Generally, however, these examples and many others suggest that the covariance assumption is reasonable. We demonstrate later that, even if the covariance assumption is violated, the consistency of LSA is not affected. Therefore, throughout the rest of the article, we assume that the covariance assumption is satisfied unless otherwise specified.

3.2. The Oracle Property

Some relevant asymptotic theories are established in this subsection. Define $a_n = \max\{\lambda_j, j \leq d_0\}$ and $b_n = \min\{\lambda_j, j > d_0\}$, and assume that the estimate $\hat{\Sigma}$ is consistent though its convergence rate can be arbitrary.

Theorem 1. (*\sqrt{n} -consistency*) *If $\sqrt{na_n} \rightarrow_p 0$, then $\hat{\beta}_{\lambda} - \beta_0 = O_p(n^{-1/2})$.*

Here “ \rightarrow_p ” stands for convergence in probability. Theorem 1 says that, as long as the maximal regularization of the relevant predictors shrinks towards 0 faster than $n^{-1/2}$, the resulting LSA estimator is \sqrt{n} -consistent (Proof in Appendix).

Next, we establish the consistency of the LSA estimator as a variable selection method. For convenience, define $\beta_a = \beta^{(Sr)} = (\beta_1, \dots, \beta_{d_0})^\top$ to be the vector corresponding to all the nonzero coefficients and $\beta_b = (\beta_{d_0+1}, \dots, \beta_d)^\top$ to be the vector corresponding to all the zero coefficients. Let $\hat{\beta}_{\lambda,a}$ and $\hat{\beta}_{\lambda,b}$ be their corresponding LSA estimators, i.e., $\hat{\beta}_{\lambda} = (\hat{\beta}_{\lambda,a}^\top, \hat{\beta}_{\lambda,b}^\top)^\top$. Similarly, write $\tilde{\beta} = (\tilde{\beta}_a^\top, \tilde{\beta}_b^\top)^\top$ and $\beta_0 = (\beta_{0a}^\top, \beta_{0b}^\top)^\top$. Then we have $\beta_{0a} \neq 0$ but $\beta_{0b} = 0$.

Theorem 2. (*Selection Consistency*) *If $\sqrt{na_n} \rightarrow_p 0$ and $\sqrt{nb_n} \rightarrow_p \infty$, then*

$$P(\hat{\beta}_{\lambda,b} = 0) \rightarrow 1.$$

Theorem 2 states that, with probability tending to one, all the zero coefficients must be estimated as 0. Theorem 1 assures the consistency of the estimators of the nonzero coefficients. Together, Theorems 1 and 2 imply that the proposed LSA estimator can identify consistently the true model (Proof in Appendix).

Theorem 3. (*Oracle Property*) *If $\sqrt{n}a_n \rightarrow 0$ and $\sqrt{nb_n} \rightarrow \infty$, then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},a} - \boldsymbol{\beta}_{0a}) \rightarrow_d N(0, \boldsymbol{\Sigma}_{S_T}).$$

Theorem 3 assures the same asymptotic distribution of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},a}$ as that of oracle. Consequently, $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},a}$ can be as efficient as oracle (Proof in Appendix).

Remark 1. Even if the covariance assumption is violated, we can verify that Theorems 1 and 2 still hold. Results similar to Theorem 3 can be established, though the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},a}$ may differ from that of oracle.

The properties established in Theorems 1, 2, and 3 rely on the appropriate specification of the tuning parameters. For a practical implementation, the problem of tuning simultaneously so many regularization parameters (i.e., λ_j) is challenging. Intuitively, some model selection criteria such as GCV, AIC, or BIC can be used to choose $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_d)^\top$ through an exhaustive grid search. However, such an exhaustive searching strategy in a d -dimensional Euclidean space is computationally demanding and practically infeasible. A simple solution is to replace each tuning parameter λ_j by the following estimate (Zou, 2006)

$$\lambda_j = \lambda_0 |\tilde{\boldsymbol{\beta}}_j|^{-\gamma}, \tag{3.4}$$

where $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1, \dots, \tilde{\boldsymbol{\beta}}_d)^\top$ is the unpenalized original estimator and $\gamma > 0$ is some pre-specified positive number. For example, $\gamma = 1$ is used in both the simulation study

and the illustration in Section 4. Such a strategy effectively transforms the originally d -dimensional tuning parameter selection problem into a univariate one. Since $\tilde{\beta}_j$ is \sqrt{n} -consistent, we can verify that the tuning parameter defined in (3.4) satisfies all the technical conditions needed by Theorems 1, 2, and 3, as long as $n^{1/2}\lambda_0 \rightarrow 0$ and $n^{(1+\gamma)/2}\lambda_0 \rightarrow \infty$. Hence, it suffices to select $\lambda_0 \in \mathbb{R}^+ = [0, \infty)$ only.

3.3. The BIC Criterion

Traditionally, GCV has been extensively used to tune regularization parameters (Tibshirani, 1996; Fan and Li, 2001). However, Wang et al. (2007) showed that the GCV approach tends to produce overfitted models, if a finite dimensional model truly exists. As a simple solution, Wang et al. (2007) developed a BIC-type selection criterion, which motivates us to consider the following

$$\text{BIC}_{\boldsymbol{\lambda}} = (\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\beta}})^{\top} \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\beta}}) + \log n \times df_{\boldsymbol{\lambda}}/n, \quad (3.5)$$

where $df_{\boldsymbol{\lambda}}$ is the number of nonzero coefficients in $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$, a simple estimate for the degrees of freedom (Zou et al., 2004).

The estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ naturally defines a candidate model $\mathcal{S}_{\boldsymbol{\lambda}} = \{j : \hat{\beta}_{\boldsymbol{\lambda},j} \neq 0\}$. Hence, one must carefully differentiate between $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ and $\hat{\boldsymbol{\beta}}_{\mathcal{S}_{\boldsymbol{\lambda}}}$, where $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ is the shrinkage estimator obtained at $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_d)^{\top}$, while $\hat{\boldsymbol{\beta}}_{\mathcal{S}_{\boldsymbol{\lambda}}}$ is the unpenalized estimator determined by (3.1) under the model identified by $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ (i.e., $\mathcal{S}_{\boldsymbol{\lambda}}$). A similar distinction also exists between $\tilde{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ and $\tilde{\boldsymbol{\beta}}_{\mathcal{S}_{\boldsymbol{\lambda}}}$. By definition (3.1), the least squares loss produced by $\hat{\boldsymbol{\beta}}_{\mathcal{S}_{\boldsymbol{\lambda}}}$ must be no larger than that of the shrinkage estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$, i.e.,

$$(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\beta}})^{\top} \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\beta}}) \geq (\hat{\boldsymbol{\beta}}_{\mathcal{S}_{\boldsymbol{\lambda}}} - \tilde{\boldsymbol{\beta}})^{\top} \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\beta}}_{\mathcal{S}_{\boldsymbol{\lambda}}} - \tilde{\boldsymbol{\beta}}). \quad (3.6)$$

According to whether the resulting model $\mathcal{S}_{\boldsymbol{\lambda}}$ is underfitted, correctly fitted, or over-

fitted, we can partition \mathbb{R}^{d^+} into the following three mutually exclusive regions

$$\begin{aligned}\mathbb{R}_-^d &= \{\boldsymbol{\lambda} \in \mathbb{R}^{d^+} : \mathcal{S}_\lambda \not\supset \mathcal{S}_T\} \\ \mathbb{R}_0^d &= \{\boldsymbol{\lambda} \in \mathbb{R}^{d^+} : \mathcal{S}_\lambda = \mathcal{S}_T\} \\ \mathbb{R}_+^d &= \{\boldsymbol{\lambda} \in \mathbb{R}^{d^+} : \mathcal{S}_\lambda \supset \mathcal{S}_T, \mathcal{S}_\lambda \neq \mathcal{S}_T\}.\end{aligned}$$

In addition, we define a reference tuning parameter sequence $\{\boldsymbol{\lambda}_n \in \mathbb{R}^d\}_{n=1}^\infty$, where the first d_0 components of $\boldsymbol{\lambda}_n$ are $1/n$ and the others are $\log n/\sqrt{n}$. Since $\boldsymbol{\lambda}_n$ satisfies the conditions in Theorems 1 and 2, it follows that $\mathcal{S}_{\boldsymbol{\lambda}_n} = \mathcal{S}_T$ with probability tending to one. We remark that $\{\boldsymbol{\lambda}_n \in \mathbb{R}^d\}_{n=1}^\infty$ is constructed only for a theoretical proof. It does not imply that the optimal tuning parameter must be $\boldsymbol{\lambda}_n$.

It can be shown that, for an arbitrary candidate model \mathcal{S} , the unpenalized estimator $\hat{\boldsymbol{\beta}}_{\mathcal{S}}$ defined in (3.1) has an explicit solution and $\hat{\boldsymbol{\beta}}_{\mathcal{S}} \rightarrow_p \boldsymbol{\beta}_{\mathcal{S}}$ for some $\boldsymbol{\beta}_{\mathcal{S}}$. Similarly, one can verify that $\boldsymbol{\beta}_{\mathcal{S}} = \boldsymbol{\beta}_0$ for any overfitted model $\mathcal{S} \supset \mathcal{S}_T$. However, for an underfitted model $\mathcal{S} \not\supset \mathcal{S}_T$, the situation is different. Because we are forcing some nonzero coefficients to be zero, we must have $\boldsymbol{\beta}_{\mathcal{S}} \neq \boldsymbol{\beta}_0$. Then, the consistency of the proposed BIC criterion (3.5) can be established.

Theorem 4. $P(\inf_{\boldsymbol{\lambda} \in \mathbb{R}_-^d \cup \mathbb{R}_+^d} BIC_\lambda > BIC_{\boldsymbol{\lambda}_n}) \rightarrow 1$.

Theorem 4 implies that any $\boldsymbol{\lambda}$ failing to identify the true model cannot be selected as the optimal tuning parameter. Therefore, the model associated with the optimal $\boldsymbol{\lambda}$ must be the true one (Proof in Appendix).

Remark 2. It can be shown that Theorem 4 holds independent of the covariance assumption. Further, if the factor $\log n$ in (3.5) is replaced by some other factor γ_n , Theorem 4 still holds as long as $\gamma_n/n \rightarrow 0$ and $\gamma_n \rightarrow \infty$.

4. NUMERICAL STUDIES

We present extensive numerical studies to demonstrate LSA’s finite sample performance. All studies are conducted in R with the *lars* package (Efron et al., 2004). This package (and many others used here) can be downloaded from the website “<http://cran.r-project.org/>”. The LSA method uses $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}$ as its standard inputs. The computational load consists of one single unpenalized full model fitting and one additional LARS processing. Therefore, the extra computational cost is minimal. All numerical studies reported in this section use the BIC criterion (3.5) for LSA (unless otherwise specified) and 500 simulation replications for each simulation setting, summarized with the median relative model error (RME), the average model size (MS), and the percentage of correct models identified (CM).

Example 1 (Logistic Regression). In this example, independent observations with binary response are generated according to the following model (Hunter and Li, 2005)

$$P(y_i = 1|\mathbf{x}_i) = \frac{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}_0\}}{1 + \exp\{\mathbf{x}_i^\top \boldsymbol{\beta}_0\}},$$

where $\boldsymbol{\beta}_0 = (3, 0, 0, 1.5, 0, 0, 2, 0, 0)^\top$. The components of \mathbf{x}_i are standard normal and the correlation between x_{ij_1} and x_{ij_2} is fixed to be $0.5^{|j_1-j_2|}$. For comparison purposes, both the LSA method and the path-finding algorithm of Park and Hastie (2006a) are used to estimate the entire solution path. Subsequently, the best solution on the path is identified by their corresponding BIC criteria. For Park and Hastie’s algorithm, the BIC criterion is defined to be $-2\log L + df_\lambda \times \log(n)$, where L stands for the log likelihood of the data. The final estimate is denoted as the PH estimator. Since the R function *glm* has both $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}$ as its standard outputs, the LSA method is implemented easily. The results are summarized in Table 1. For this example, the

LSA estimator slightly outperforms the PH estimator.

Table 1: Simulation Results for Logistic Regression

Sample Size	Estimation Method	RME		MS		CM	
		Median	(SE)	Mean	(SE)	Mean	(SE)
200	LSA	0.504	(0.015)	3.178	(0.026)	0.798	(0.018)
	PH	0.544	(0.017)	3.272	(0.033)	0.716	(0.020)
400	LSA	0.417	(0.013)	3.130	(0.018)	0.888	(0.014)
	PH	0.566	(0.024)	3.092	(0.023)	0.846	(0.016)

Example 2 (Cox’s Model). In this simulation study, independent survival data are generated according to the following hazard function (Fan and Li, 2002)

$$h(t_i|\mathbf{x}_i) = \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_0), \quad (4.1)$$

where t_i is the survival time from the i th subject and $\boldsymbol{\beta}_0 = (0.8, 0, 0, 1, 0, 0, 0.6, 0)^\top$. According to (4.1), the survival time of the i th subject is exponentially distributed with mean $\exp(-\mathbf{x}_i^\top \boldsymbol{\beta}_0)$. Again, the covariate x_{ij} is generated in the same manner as the previous example. Furthermore, independent censoring time is generated from an exponential distribution with mean $u \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_0)$, where u is a random variable uniformly distributed on $[1, 3]$. Such a censoring mechanism produces about 30% censored data (Fan and Li, 2002). With the function *coxph* in the R package *survival*, the LSA method is implemented easily. As before, we compare the performance of the LSA estimator to that of the PH estimator, using for the latter the BIC criterion $-2 \log L + df_\lambda \log(n)$, where L is the partial likelihood (Cox, 1972). The simulation results, summarized in Table 2, confirm similar performance in the LSA and PH estimators.

Example 3 (LAD Regression). As our third example, we revisit the LAD-LASSO

Table 2: Simulation Results for Cox’s Model

Sample Size	Estimation Method	RME		MS		CM	
		Median	(SE)	Mean	(SE)	Mean	(SE)
100	LSA	0.440	(0.038)	3.394	(0.031)	0.644	(0.021)
	PH	0.451	(0.047)	3.342	(0.033)	0.634	(0.022)
200	LSA	0.575	(0.065)	3.208	(0.022)	0.822	(0.017)
	PH	0.643	(0.096)	3.182	(0.022)	0.824	(0.017)

problem studied by Wang et al. (2007). Data are simulated according to

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta}_0 + \sigma \times e_i,$$

where we take $\sigma = 1$, \mathbf{x}_i is generated in the same manner as in Example 1, and $\boldsymbol{\beta}_0 = (3, 1.5, 0, 0, 2, 0, 0, 0)^\top$. Furthermore, e_i is simulated from a mixture distribution with 90% of the observations coming from a standard normal distribution and 10% from a standard Cauchy distribution. As Wang et al. (2007) did not provide a simple path finding algorithm, we select the optimal λ_0 through a grid search from 50 equally spaced λ_0 values. We refer to the corresponding estimator as the WLJ estimator. Using the function *rq* in the R package *quantreg*, the LSA method is implemented easily. The BIC criterion used by the WLJ estimator is given in Hurvich and Tsai (1990, Eq. 8). The results, summarized in Table 3, indicate a slightly better performance of the LSA estimator over the WLJ estimator.

Example 4 (Tukey’s Biweight Regression). In this example, we consider Tukey’s biweight regression (Venables and Ripley, 1999). Data are simulated in the same manner as the previous example but with 80% of observations from a standard normal distribution and 20% from a standard Cauchy distribution. The unpenalized estimator

Table 3: Simulation Results for LAD Regression

Sample Size	Estimation Method	RME		MS		CM	
		Median	(SE)	Mean	(SE)	Mean	(SE)
100	LSA	0.463	(0.017)	3.022	(0.007)	0.980	(0.006)
	WLJ	0.516	(0.018)	3.262	(0.028)	0.790	(0.018)
200	LSA	0.429	(0.013)	3.010	(0.004)	0.990	(0.004)
	WLJ	0.453	(0.015)	3.148	(0.021)	0.876	(0.015)

$\tilde{\beta}$ is obtained by minimizing the following objective function

$$\sum_{i=1}^n \rho(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}),$$

where $\rho(\cdot)$ is Tukey's biweight function (Venables and Ripley, 1999). Tukey's biweight regression is a special case of M-estimation. The covariance assumption is satisfied as long as the residual is independent of the predictor (Huber, 1981, Section 6.3). Using the function *rlm* in the R package *MASS*, the LSA method is implemented easily. The results, summarized in Table 4, confirm LSA's good performance.

Table 4: Simulation Results for Tukey's Biweight Regression

Sample Size	RME		MS		CM	
	Median	(SE)	Mean	(SE)	Mean	(SE)
50	0.528	(0.015)	3.342	(0.030)	0.750	(0.019)
100	0.495	(0.015)	3.160	(0.020)	0.864	(0.015)
200	0.455	(0.015)	3.076	(0.013)	0.928	(0.012)

Example 5 (Buckley-James Estimator). As our fifth simulation example, we consider the Buckley-James estimator (Buckley and James, 1979), an attractive alternative to Cox's proportional hazards model. Survival data are generated according to the following log-linear model

$$\log(t_i) = \mathbf{x}_i^\top \boldsymbol{\beta}_0 + \sigma \times e_i,$$

where t_i is the survival time from the i th subject, $\beta_0 = (0.8, -0.5, 0, 0, 0.5, 0, 0, 0)^\top$, $\sigma = 0.5$, and the covariate \mathbf{x}_i is generated in the same manner as in Example 1. To create censored observations, random censoring times are generated from a uniform distribution on $[0, 4]$, which produces about 40% censored observations. Using the R function *bj* in the package *Design*, the proposed LSA method is implemented easily. The asymptotic covariance matrix of the Buckley-James estimator is of the form $\Sigma = \mathbf{W}^{-1}\mathbf{V}\mathbf{W}^{-1}$ for some matrix \mathbf{W} and \mathbf{V} (Lai and Ying, 1991). Therefore, the covariance assumption can be violated. Nevertheless, Table 5 shows that the LSA estimator selects the true model consistently and estimates the parameters efficiently.

Table 5: Simulation Results for Buckley-James Estimator

Sample Size	RME		MS		CM	
	Median	(SE)	Mean	(SE)	Mean	(SE)
50	0.697	(0.021)	3.326	(0.030)	0.712	(0.020)
100	0.582	(0.020)	3.166	(0.020)	0.860	(0.016)
200	0.553	(0.014)	3.108	(0.015)	0.900	(0.013)

Example 6 (Confidence Intervals). As pointed out by one anonymous referee, the oracle property implies that asymptotically valid confidence intervals can be constructed for the nonzero coefficients. We make no attempt to construct confidence intervals for the zero coefficients, because they are expected to be estimated as 0. For illustration purposes, Example 2 is revisited with the same tuning parameter selection scheme. For a given LSA estimator and its nonzero components, we estimate its asymptotic covariance matrix by inverting the corresponding submatrix of $\hat{\Sigma}^{-1}$. An asymptotically valid 95% confidence interval is then constructed via the normal approximation. The actual coverage probabilities (in percentage) are reported in Table 6, along with those of adaptive LASSO (aLASSO) and oracle (ORACLE).

For relatively large sample sizes (e.g., $n = 200$), the performance of LSA is com-

parable to that of oracle, which further confirms our asymptotic theory. With smaller sample sizes (e.g., $n = 100$), the performance of LSA becomes slightly worse but is still practically acceptable. The decreased performance may not be due to the least squares approximation, but rather may be inherent in the adaptive LASSO penalty: results for LSA and aLASSO are very similar. Experiments are also conducted for Examples 1, 3, and 4 with similar findings.

Table 6: Confidence Interval Coverage Probability for Cox’s Model

Estimation Method	Sample Size					
	$n = 100$			$n = 200$		
	β_{01}	β_{04}	β_{07}	β_{01}	β_{04}	β_{07}
LSA	92.4	95.4	93.2	94.8	96.8	96.8
aLASSO	91.4	93.2	92.2	91.4	93.0	94.2
ORACLE	95.2	94.4	95.2	93.8	94.6	96.8

Example 7 (South Africa Heart Disease Data). As our first real data example, we consider the South Africa heart disease data in Park and Hastie (2006a). The goal of this study is to establish the intensity of ischemic heart risk factors in three rural areas of the West Cape, South Africa. For this dataset, there are 9 predictors associated with a binary response variable, which indicates the presence or absence of myocardial infarction for 462 subjects. To compare the differences among the original LASSO, the adaptive LASSO, and the proposed LSA estimator, we plot the solution paths of these three estimates in the left panel of Figure 1. For the PH estimates, we set the arc length to be 0.1 (Park and Hastie, 2006a). Little difference is observed between the the adaptive PH estimate and our LSA estimate. Both give a model with the same set of nonzero coefficients. For this example, the same tuning parameter selection scheme as in Example 1 is used.

Example 8 (Primary Biliary Cirrhosis Data). As our second illustration, we revisit

the Primary Biliary Cirrhosis Data, which was collected by the Mayo Clinic between January 1974 and May 1984. The data include 312 randomized patients having primary biliary cirrhosis (PBC) with 18 covariates and 276 complete records (Tibshirani, 1997). The data are analyzed in a similar manner to the previous example and the results are presented in the right panel of Figure 1. Once again, little difference is observed between the LSA and the adaptive PH estimates. Furthermore, both give a model with the same set of nonzero coefficients. For this example, the same tuning parameter selection scheme as in Example 1 is used.

Example 9 (Leukemia Microarray Data). As noted by one anonymous referee, it would be more attractive if the proposed LSA method can be potentially useful in the $d \gg n$ situation. Consequently, we devote this example to evaluate this possibility via the leukemia microarray dataset (Golub et al., 1999). A total of 72 samples are contained in the dataset. The response of interest is the presence of acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML). The predictors associated with each sample are the expression levels measured on 3571 genes (i.e., $d = 3571$); see Dudoit et al. (2002). As in Golub et al. (1999), the first 38 (i.e., $n = 38$) samples are used for training, while the remaining 34 samples are reserved for testing. Clearly, the predictor dimension ($d = 3571$) is substantially larger than the sample size ($n = 38$). Hence, LSA cannot be applied directly.

To resolve this difficulty, we make use of the L_2 -penalized logistic regression recently developed by Park and Hastie (2006b), and replace the inputs $(\tilde{\beta}, \hat{\Sigma})$ by their corresponding regularized estimates $(\tilde{\beta}_\kappa, \hat{\Sigma}_\kappa)$, where $\kappa > 0$ is the tuning parameter used by the L_2 penalty. To avoid the double shrinkage effect (Zou and Hastie, 2005), we set κ to be some small positive number (e.g., $\kappa = 10^{-3}$ for this example). LSA is then used to find the solution path, from which the best solution is identified by 10-fold

cross-validation. Our experience suggests that the BIC criterion (3.5) is not applicable here due to the excessively large dimensionality. We subsequently use the estimated parameters to predict the class memberships of the 34 testing samples. LSA identifies 54 important genes with zero out-of-sample prediction error.

As a cautionary note, we consider this example as only a very preliminary evaluation about the potential usefulness of the LSA method in the $d \gg n$ situation. Theoretically, we are not clear about whether the covariance assumption is satisfied here. Computationally, we find that the small sample performance of the 10-fold cross-validation method could be unstable. Despite our encouraging findings, further research is definitely needed.

5. CONCLUDING REMARKS

We conclude the article with two remarks. First, the model selection consistency of the LSA method depends on large sample theory. However, as noted by one anonymous referee, for a given dataset with a fixed sample size n and an unknown parameter β_0 , it is difficult to assess whether the sample size is sufficient. Hence, it is of interest to evaluate the finite sample performance of the LSA method under some local perturbation. For this purpose, we follow the idea of Leeb and Pötscher (2005) and replicate Example 1 with regression coefficients replaced by $(3, 0, 0, 1.5, 0, 0, 2, 0, 0)/\sqrt{n}$. The simulation results are summarized in Table 7 in a manner similar to Example 6. The results show that both LSA and adaptive LASSO (aLASSO) perform poorly. These results further confirm the theoretical findings of Leeb and Pötscher (2005), which states that no consistent model selection method including LSA can have a good minimax efficiency. Then, how to reconcile the conflict between selection consistency and minimax efficiency becomes a problem of interest.

Table 7: Simulation Results for Logistic Regression under Local Perturbation

Sample Size	Estimation Method	Coverage Probability of 95% Confidence Interval in %			% of Correct Models
		β_1	β_4	β_7	
200	LSA	9.2	3.8	4.2	1.0
	aLASSO	9.0	3.8	4.2	0.8
	ORACLE	94.8	95.6	95.2	100.0
400	LSA	11.4	3.8	5.0	1.6
	aLASSO	11.2	3.8	5.0	1.6
	ORACLE	95.8	94.2	94.8	100.0

Secondly, the covariance assumption can be violated (see for example, Example 5 in Section 5). However, even if the covariance assumption is violated, the LSA method is still consistent for both parameter estimation and model selection. How to further improve LSA's efficiency under the covariance assumption violation is another interesting topic for future study.

APPENDIX

Proof of Theorem 1: Note that the LSA objective function $Q(\beta)$ given in (2.5) is a strictly convex function in β . Hence, as long as we can show that (2.5) has a \sqrt{n} -consistent local minimizer, it must be \sqrt{n} -consistent global minimizer. Thus, the theorem's conclusion follows immediately. As in Fan and Li (2001), the existence of a \sqrt{n} -consistent local minimizer is implied by that fact that for an arbitrarily small $\epsilon > 0$, there exists a sufficiently large constant C , such that

$$\liminf_n P \left\{ \inf_{u \in \mathbb{R}^d: \|u\|=C} Q(\beta_0 + n^{-1/2}u) > Q(\beta_0) \right\} > 1 - \epsilon. \quad (\text{A.1})$$

Let $\mathbf{u} = (u_1, \dots, u_d)^\top$. Then, by the definition of (2.5), simple algebra shows that

$$\begin{aligned}
& n \left\{ Q(\boldsymbol{\beta}_0 + n^{-1/2} \mathbf{u}) - Q(\boldsymbol{\beta}_0) \right\} \\
&= \mathbf{u}^\top \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{u} + 2\mathbf{u}^\top \hat{\boldsymbol{\Sigma}}^{-1} \left[\sqrt{n}(\boldsymbol{\beta}_0 - \tilde{\boldsymbol{\beta}}) \right] + n \sum_{j=1}^d \lambda_j |\boldsymbol{\beta}_{0j} + n^{-1/2} u_j| - n \sum_{j=1}^d \lambda_j |\boldsymbol{\beta}_{0j}| \\
&= \mathbf{u}^\top \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{u} + 2\mathbf{u}^\top \hat{\boldsymbol{\Sigma}}^{-1} \left[\sqrt{n}(\boldsymbol{\beta}_0 - \tilde{\boldsymbol{\beta}}) \right] + n \sum_{j=1}^{d_0} \lambda_j |\boldsymbol{\beta}_{0j} + n^{-1/2} u_j| - n \sum_{j=1}^{d_0} \lambda_j |\boldsymbol{\beta}_{0j}| \\
&\geq \mathbf{u}^\top \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{u} + 2\mathbf{u}^\top \hat{\boldsymbol{\Sigma}}^{-1} \left[\sqrt{n}(\boldsymbol{\beta}_0 - \tilde{\boldsymbol{\beta}}) \right] + n \sum_{j=1}^{d_0} \lambda_j (|\boldsymbol{\beta}_{0j} + n^{-1/2} u_j| - |\boldsymbol{\beta}_{0j}|) \\
&\geq \mathbf{u}^\top \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{u} + 2\mathbf{u}^\top \hat{\boldsymbol{\Sigma}}^{-1} \left[\sqrt{n}(\boldsymbol{\beta}_0 - \tilde{\boldsymbol{\beta}}) \right] - d_0 (\sqrt{n} a_n) \|\mathbf{u}\|, \tag{A.2}
\end{aligned}$$

where the second equality holds because $\boldsymbol{\beta}_{0j} = 0$ for any $j > d_0$ by the model assumption. In addition to that, according to the theorem's condition $\sqrt{n} a_n \rightarrow_p 0$, we know that the third term in (A.2) is $o_p(1)$. Furthermore, because $\|\mathbf{u}\| = C$, the first term in (A.2) is uniformly larger than $\tau_{\min}(\hat{\boldsymbol{\Sigma}}^{-1}) C^2 \rightarrow_p \tau_{\min}(\boldsymbol{\Sigma}^{-1}) C^2$, where $\tau_{\min}(M)$ refers to the minimal eigenvalue of M . It follows then, with probability tending to one, that the first term in (A.2) is uniformly larger than $0.5 \tau_{\min}(\boldsymbol{\Sigma}^{-1}) C^2$, which is quadratic in C . On the other hand, the second term in (A.2) is uniformly bounded by $C \|\hat{\boldsymbol{\Sigma}}^{-1} \sqrt{n}(\boldsymbol{\beta}_0 - \tilde{\boldsymbol{\beta}})\|$, which is linear in C with coefficient $\|\hat{\boldsymbol{\Sigma}}^{-1} \sqrt{n}(\boldsymbol{\beta}_0 - \tilde{\boldsymbol{\beta}})\| = O_p(1)$. Therefore, as long as the constant C is sufficiently large, the first term will always dominate the other two with arbitrarily large probability. This implies the inequality (A.1), and the proof is completed.

Proof of Theorem 2: We need only show that $P(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},j} = 0) \rightarrow 1$ for any $d_0 < j \leq d$.

If $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},j} \neq 0$ for some $d_0 < j \leq d$,

$$\sqrt{n} \times \frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j} \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}} = 2\hat{\boldsymbol{\Omega}}^{(j)} \times \sqrt{n}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\beta}}) + \sqrt{n} \lambda_j \text{sgn}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},j}), \tag{A.3}$$

where $\hat{\Omega}^{(j)}$ stands for the j th row of $\hat{\Omega} = \hat{\Sigma}^{-1}$. Because $\hat{\Omega} \rightarrow_p \Omega = \Sigma^{-1}$ and $\sqrt{n}(\hat{\beta}_\lambda - \tilde{\beta}) = O_p(1)$, the first term in (A.3) is $O_p(1)$. On other hand, recall that j is some index satisfying $d_0 < j \leq d$, so that by the theorem's condition we know that $\sqrt{n}\lambda_j \geq \sqrt{nb_n} \rightarrow \infty$. As a result, with probability tending to one, either $\hat{\beta}_{\lambda,j} = 0$ or the sign of (A.3) is equal to the sign of $\hat{\beta}_{\lambda,j}$. But the latter possibility would imply that $\hat{\beta}_\lambda$ is not the minimizer of Q , whereas, from the proof of Theorem 1, the minimizer does exist, again with probability tending to one. Taken together these facts imply that $P(\hat{\beta}_{\lambda,j} = 0) \rightarrow 1$. This completes the proof.

Proof of Theorem 3: For convenience purposes, first decompose the asymptotic covariance matrix Σ into the following block matrix form

$$\Sigma = \Sigma_{S_F} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

where Σ_{11} is the first $d_0 \times d_0$ submatrix. It follows then that $\Sigma_{11} = \Sigma^{(S_T)}$. Similarly, its inverse matrix $\Omega = \Omega_{S_F} = \Sigma^{-1}$ can also be partitioned as

$$\Sigma^{-1} = \Omega_{S_F} = \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix}.$$

The estimators $(\hat{\Sigma}_{ij}, \hat{\Omega}_{ij})$ ($i, j \in \{1, 2\}$) can be similarly defined. According to Theorem 2, with probability tending to one, $\hat{\beta}_{\lambda,b} = 0$, hence, $\hat{\beta}_{\lambda,a}$ must be the global minimizer of the objective function

$$Q_0(\beta_a) = (\beta_a - \tilde{\beta}_a)^\top \hat{\Omega}_{11}(\beta_a - \tilde{\beta}_a) - 2(\beta_a - \tilde{\beta}_a)^\top \hat{\Omega}_{12}\tilde{\beta}_b + \tilde{\beta}_b^\top \hat{\Omega}_{22}\tilde{\beta}_b + \sum_{j=1}^{d_0} \lambda_j |\beta_j|.$$

By Theorem 1, with probability tending to 1, each component of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},a}$ must be nonzero, so that the partial derivative $\partial Q(\boldsymbol{\beta})/\partial \boldsymbol{\beta}_j$ exists at $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ for $1 \leq j \leq d_0$. Hence, the following normal equation must be satisfied

$$0 = \frac{1}{2} \times \frac{\partial Q_0(\boldsymbol{\beta}_a)}{\partial \boldsymbol{\beta}_a} \Big|_{\boldsymbol{\beta}_a = \hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},a}} = \hat{\boldsymbol{\Omega}}_{11}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},a} - \tilde{\boldsymbol{\beta}}_a) - \hat{\boldsymbol{\Omega}}_{12}\tilde{\boldsymbol{\beta}}_b + D(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},a}), \quad (\text{A.4})$$

where $D(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},a})$ is a d_0 -dimensional vector with its j th component given by $0.5\lambda_j \text{sgn}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},j})$.

Note that (A.4) implies that

$$\begin{aligned} \sqrt{n}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},a} - \boldsymbol{\beta}_{0a}) &= \sqrt{n}(\tilde{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_{0a}) + \hat{\boldsymbol{\Omega}}_{11}^{-1}\hat{\boldsymbol{\Omega}}_{12}(\sqrt{n}\tilde{\boldsymbol{\beta}}_b) - \hat{\boldsymbol{\Omega}}_{11}^{-1} \times \sqrt{n}D(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},a}) \\ &= \sqrt{n}(\tilde{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_{0a}) + \hat{\boldsymbol{\Omega}}_{11}^{-1}\hat{\boldsymbol{\Omega}}_{12}(\sqrt{n}\tilde{\boldsymbol{\beta}}_b) + o_p(1) \end{aligned} \quad (\text{A.5})$$

$$= \sqrt{n}(\tilde{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_{0a}) + \boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}(\sqrt{n}\tilde{\boldsymbol{\beta}}_b) + o_p(1), \quad (\text{A.6})$$

where the equality (A.5) is due mainly to the fact that $\sqrt{n}D(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},a}) = o_p(1)$ because $\sqrt{n}\lambda_j < \sqrt{n}a_n \rightarrow_p 0$ for any $j \leq d_0$, and the equality (A.6) is due to $\sqrt{n}\tilde{\boldsymbol{\beta}}_b = O_p(1)$.

Furthermore, because $\boldsymbol{\Omega}_{\mathcal{S}_F} = \boldsymbol{\Sigma}_{\mathcal{S}_F}^{-1}$, it can be verified that

$$\boldsymbol{\Omega}_{\mathcal{S}_F} = \begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{(11)}^{-1} & -\boldsymbol{\Sigma}_{(11)}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{(11)}^{-1} & \boldsymbol{\Sigma}_{22}^{-1} + \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{(11)}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \end{bmatrix},$$

where $\boldsymbol{\Sigma}_{(11)} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. Then, it follows that $\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12} = -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$. Plugging this equality back to (A.6), we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},a} - \boldsymbol{\beta}_{0a}) = \sqrt{n}(\tilde{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_{0a}) - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\sqrt{n}\tilde{\boldsymbol{\beta}}_b) + o_p(1),$$

which is asymptotically normal with mean 0 and inverse asymptotic covariance matrix

$$\Sigma_{(11)}^{-1} = \left(\Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \right)^{-1} = \Omega_{11}.$$

Under the covariance assumption, we have $\Omega_{11} = \Omega_{\mathcal{S}_T}$, which implies the same asymptotic covariance matrices between the LSA estimator $\hat{\beta}_{\lambda,a}$ and the oracle estimator $\tilde{\beta}_{\mathcal{S}_T}^{(\mathcal{S}_T)}$. This further implies that the final estimator $\hat{\beta}_{\lambda,a}$ shares the same asymptotic distribution, hence, efficiency as the oracle estimator $\tilde{\beta}_{\mathcal{S}_T}^{(\mathcal{S}_T)}$. Therefore, the oracle property is established and the proof is completed.

Proof of Theorem 4: Following a similar idea in Wang et al. (2007), for any $\mathcal{S}_\lambda \neq \mathcal{S}_T$, we consider two different cases according to whether the model is underfitted or overfitted. In either case, we show that the theorem conclusion is valid. Specifically,

Case 1 (Underfitted Model). First, note that λ_n satisfies the regularity conditions as specified by Theorem 1. Consequently, the resulting estimator $\hat{\beta}_{\lambda_n}$ is \sqrt{n} -consistent, which immediately implies that its associated BIC value is of the order $o_p(1)$. On the other hand, due to the fact that $\beta_{\mathcal{S}} \neq \beta_0$ for any $\mathcal{S} \not\supseteq \mathcal{S}_T$, we know that

$$\begin{aligned} \text{BIC}_\lambda &= (\hat{\beta}_\lambda - \tilde{\beta})^\top \hat{\Sigma}^{-1} (\hat{\beta}_\lambda - \tilde{\beta}) + df_\lambda \times \log(n)/n \\ &\geq (\hat{\beta}_\lambda - \tilde{\beta})^\top \hat{\Sigma}^{-1} (\hat{\beta}_\lambda - \tilde{\beta}) \geq (\hat{\beta}_{\mathcal{S}_\lambda} - \tilde{\beta})^\top \hat{\Sigma}^{-1} (\hat{\beta}_{\mathcal{S}_\lambda} - \tilde{\beta}) \quad (\text{A.7}) \\ &\geq \min_{\mathcal{S} \not\supseteq \mathcal{S}_T} (\hat{\beta}_{\mathcal{S}} - \tilde{\beta})^\top \hat{\Sigma}^{-1} (\hat{\beta}_{\mathcal{S}} - \tilde{\beta}) \rightarrow_p \min_{\mathcal{S} \not\supseteq \mathcal{S}_T} (\beta_{\mathcal{S}} - \beta)^\top \Sigma^{-1} (\beta_{\mathcal{S}} - \beta) > 0 \quad (\text{A.8}) \end{aligned}$$

where the inequality (A.7) is due to (3.6) and the last inequality in (A.8) is due to the fact that Σ is positive definite and $\beta_{\mathcal{S}} \neq \beta_0$ for any $\mathcal{S} \not\supseteq \mathcal{S}_T$. Consequently, with probability tending to one, we must have $\inf_{\lambda \in \mathbb{R}_+^d} \text{BIC}_\lambda > \text{BIC}_{\lambda_n}$. Hence, it follows immediately that $P(\inf_{\lambda \in \mathbb{R}_+^d} \text{BIC}_\lambda > \text{BIC}_{\lambda_n}) \rightarrow 1$.

Case 2 (Overfitted Model). Let $\boldsymbol{\lambda}$ be an arbitrary tuning parameter, which produces an overfitted model (i.e., $\boldsymbol{\lambda} \in \mathbb{R}_+^d$). We then have $df_{\boldsymbol{\lambda}} - d_0 \geq 1$ and

$$\begin{aligned}
n(\text{BIC}_{\boldsymbol{\lambda}} - \text{BIC}_{\boldsymbol{\lambda}_n}) &= n(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\beta}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\beta}}) - n(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_n} - \tilde{\boldsymbol{\beta}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_n} - \tilde{\boldsymbol{\beta}}) + (df_{\boldsymbol{\lambda}} - d_0) \log n \\
&\geq n(\hat{\boldsymbol{\beta}}_{\mathcal{S}_{\boldsymbol{\lambda}}} - \tilde{\boldsymbol{\beta}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\beta}}_{\mathcal{S}_{\boldsymbol{\lambda}}} - \tilde{\boldsymbol{\beta}}) - n(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_n} - \tilde{\boldsymbol{\beta}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_n} - \tilde{\boldsymbol{\beta}}) + (df_{\boldsymbol{\lambda}} - d_0) \log n \\
&\geq n(\hat{\boldsymbol{\beta}}_{\mathcal{S}_{\boldsymbol{\lambda}}} - \tilde{\boldsymbol{\beta}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\beta}}_{\mathcal{S}_{\boldsymbol{\lambda}}} - \tilde{\boldsymbol{\beta}}) - n(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_n} - \tilde{\boldsymbol{\beta}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_n} - \tilde{\boldsymbol{\beta}}) + \log n \\
&\geq \inf_{\mathcal{S} \supset \mathcal{S}_T} n(\hat{\boldsymbol{\beta}}_{\mathcal{S}} - \tilde{\boldsymbol{\beta}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\beta}}_{\mathcal{S}} - \tilde{\boldsymbol{\beta}}) - n(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_n} - \tilde{\boldsymbol{\beta}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_n} - \tilde{\boldsymbol{\beta}}) + \log n, \quad (\text{A.9})
\end{aligned}$$

where the first term in (A.9) is $O_p(1)$ because $\hat{\boldsymbol{\beta}}_{\mathcal{S}}$ is \sqrt{n} -consistent for any overfitted model $\mathcal{S} \supset \mathcal{S}_T$ according to (3.2). On the other hand, its second term is also $O_p(1)$ because $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_n}$ is \sqrt{n} -consistent according to Theorem 1. Lastly, note that the last term in (A.9) diverges to $+\infty$ as $n \rightarrow \infty$. As a result, we must have $P(\inf_{\boldsymbol{\lambda} \in \mathbb{R}_+^d} \text{BIC}_{\boldsymbol{\lambda}} > \text{BIC}_{\boldsymbol{\lambda}_n}) \rightarrow 1$. This finishes case 2. Hence, the whole proof is completed.

REFERENCES

- Buckley, J. and James, I. (1979), “Linear regression with censored data,” *Biometrika*, 66, 429–36.
- Cox, D. R. (1972), “Regression models and life tables,” *Journal of the Royal Statistical Society, Series B*, 34, 187–220.
- Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). “Comparison of discrimination methods for the classification of tumors using gene expression data,” *Journal of the American Statistical Association*, 97, 77–87.
- 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.
- (2002), “Variable selection for Cox’s proportional hazards model and frailty model,” *The Annals of Statistics*, 30, 74–99.
- Golub TR, Slonim DK, Tamayo P, Huard, C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. (1999), “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring,” *Science*, 286(5439), 531–7.
- Huber, P. (1981), *Robust Estimation*, John Wiley, New York, NY.
- Hunter, D. R. and Li, R. (2005), “Variable selection using MM algorithms,” *The Annals of Statistics*, 33, 1617–1642.
- Hurvich, C. M. and Tsai, C. L. (1990), “Model selection for least absolute deviation regression in small samples,” *Statistics and Probability Letters*, 9, 259–265.
- Knight, K. and Fu, W. (2000), “Asymptotics for lasso-type estimators,” *The Annals of Statistics*, 28, 1356–1378.
- Koenker, R. and Bassett, G. (1978), “Regression quantiles,” *Econometrica*, 46, 33–50.
- Lai, T. L. and Ying, Z. (1991), “Large sample theory for a modified Buckley-James estimator for regression analysis with censored data,” *The Annals of Statistics*, 19, 1370–1402.
- Leeb, H. and Pötscher, B. M. (2005), “Sparse estimators and the oracle property, or the return of Hodges’ estimator,” *Technical Report, Department of Statistics, Yale University*.

- Leeb, H. and Pötscher, B. M. (2006), “Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk-bound results,” *Econometric Theory*, 22, 69–97.
- Leng, C., Lin, Y., and Wahba, G. (2006), “A note on lasso and related procedures in model selection,” *Statistica Sinica*, 16(4), 1273–1284.
- McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, Chapman and Hall, New York, NY.
- Osborne, M.R., Presnell, B., and Turlach, B. A. (2000), “A new approach to variable selection in least squares problems,” *IMA Journal of Numerical Analysis*, 20(3), 389–404.
- Park, M. Y. and Hastie, T. (2006a), “An L1 regularization-path algorithm for generalized linear models,” *Manuscript, Department of Statistics, Stanford University*.
- Park, M. Y. and Hastie, T. (2006b), “Penalized logistic regression for detecting gene interactions,” *Manuscript, Department of Statistics, Stanford University*.
- Pollard, D. (1991), “Asymptotics for least absolute deviation regression estimators,” *Econometric Theory*, 7, 186–199.
- Rosset, S. (2004), “Tracking curved regularized optimization solution paths,” *NIPS 2004*.
- Tibshirani, R. J. (1996), “Regression shrinkage and selection via the LASSO,” *Journal of the Royal Statistical Society, Series B*, 58, 267–288.
- Tibshirani, R. J. (1997), “The LASSO method for variable selection in the Cox model,” *Statistics in Medicine*, 16, 385–395.

- Venables, W. N. and Ripley, B. D. (1999), *Modern Applied Statistics with S-Plus, 3rd ed.*, New York: Springer.
- Wang, H., Li, G., and Jiang, G. (2007), “Robust regression shrinkage and consistent variable selection via the LAD-LASSO,” *Journal of Business and Economics Statistics*, To appear.
- Wang, H., Li, R., and Tsai, C. L. (2007), “On the Consistency of SCAD Tuning Parameter Selector,” *Biometrika*, To appear.
- Yuan, M. and Lin, Y. (2007), “On the nonnegative garrote estimator,” *Journal of the Royal Statistical Society, Series B*. To appear.
- Zhao, P. and Yu, B. (2004), “Boosted lasso,” *Technical Report #678, Statistics, UC Berkeley*.
- Zou, H. (2006), “The adaptive lasso and its oracle properties,” *Journal of the American Statistical Association*, 101, 1418–1429.
- Zou, H. and Hastie, T. (2006), “Regression shrinkage and selection via the elastic net with application to microarrays”, *Journal of the Royal Statistical Society, Series B*, 67, 301–320.
- Zou, H., Hastie, T., and Tibshirani, R. (2004), “On the ‘degrees of freedom’ of lasso,” *Technical Report, Statistics Department, Stanford University*.

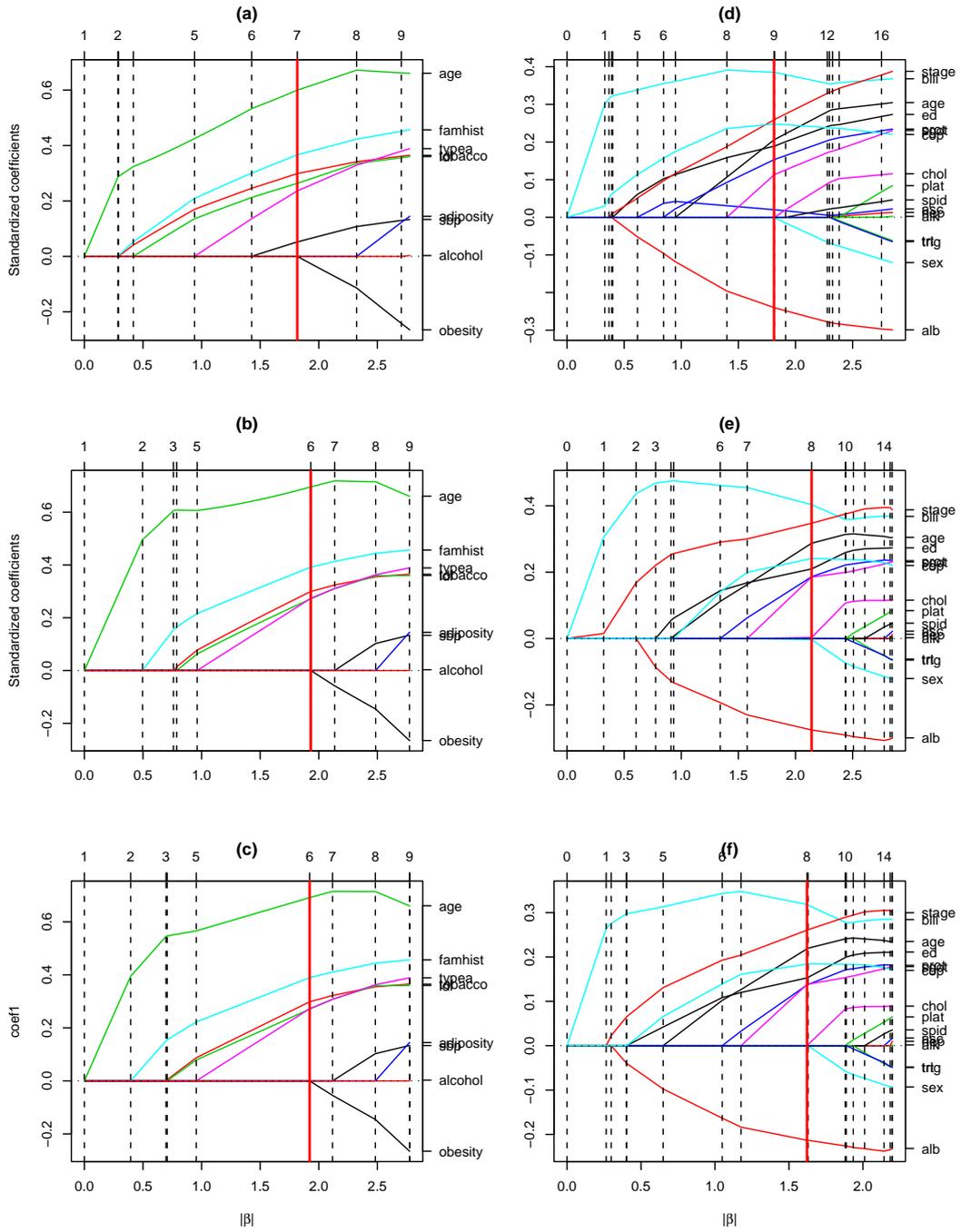


Figure 1: Left Panel: South Africa Heart Disease Data; Right Panel: Primary Biliary Cirrhosis Data; (a) and (d): the solution paths for the original LASSO via PH algorithm; (b) and (e): the solution paths for the adaptive LASSO via PH algorithm; (c) and (f): the solution paths via LSA. The vertical dashed lines indicate where the active sets are modified, and the vertical solid lines indicate the best fit according to BIC.