Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Unified LASSO Estimation by Least Squares Approximation

Hansheng Wang & Chenlei Leng

To cite this article: Hansheng Wang & Chenlei Leng (2007) Unified LASSO Estimation by Least Squares Approximation, Journal of the American Statistical Association, 102:479, 1039-1048, DOI: 10.1198/016214507000000509

To link to this article: https://doi.org/10.1198/016214507000000509

Unified LASSO Estimation by Least Squares Approximation

Hansheng WANG and Chenlei LENG

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 Bayes information criterion–type tuning parameter selector are used, the resulting LSA estimator can be as efficient as the oracle. Extensive numerical studies confirm our theory.

KEY WORDS: Adaptive LASSO; Bayes information criterion; LASSO; Least angle regression; Least squares approximation; Microarray data; Oracle Property; Solution path.

1. INTRODUCTION

The least absolute shrinkage and selection operator (LASSO; Tibshirani 1996) is a useful and well-studied approach to the problem of variable selection (Knight and Fu 2000; Fan and Li 2001; Leng, Lin, and Wahba 2006; Wang et al. 2007a,b; 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 that is as efficient as the oracle (Zou 2006; Wang et al. 2007a,b). Furthermore, LASSO has very good computational properties. In the ordinary least squares (OLS) setting, Osborne, Presnell, and Turlach (2000) and Efron, Hastie, Johnstone, and Tibshirani (2004) have shown 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 also have been 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 proven very useful in both theory and practice. Furthermore, with reasonable programming effort, these algorithms can be easily modified 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. Consequently, 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, such as quantile regression (Koenker and Bassett 1978). Theoretically, similar asymptotic and numerical theories can be established for these models in a case-by-case manner. But 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 OLS, GLMs, 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. 2007a,b) and a Bayes information criterion (BIC)-type tuning parameter selector (Wang et al. 2007b) are used, the resulting LSA estimator can be as efficient as the oracle.

The rest of the article is organized as follows. Section 2 introduces the LSA method, and Section 3 discusses this method's main theoretical properties. Section 4 presents extensive numerical studies, and Section 5 concludes the article with some discussion. The Appendix contains the technical details.

2. THE LEAST SQUARES APPROXIMATION

2.1 Model and Notation

Let $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ be n independent and identically distributed random vectors, where $y_i \in \mathbb{R}^1$ is the response of interest and $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})^\top \in \mathbb{R}^d$ is the d-dimensional predictor. Assume that $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^\top \in \mathbb{R}^d$ 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, then $\tilde{\boldsymbol{\beta}}$ is the usual OLS estimate; If $\mathcal{L}_n(\boldsymbol{\beta})$ is the negative log-likelihood function, then $\tilde{\boldsymbol{\beta}}$ 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 $\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathrm{N}(0, \boldsymbol{\Sigma})$ for some true

Hansheng Wang is Associate Professor, Guanghua School of Management, Peking University, Beijing, P. R. China, 100871 (E-mail: *hansheng@gsm.pku.edu.cn*). Chenlei Leng is Assistant Professor, Department of Statistics and Applied Probability, National University of Singapore, Sinagapore (E-mail: *stalc@nus.edu.sg*). Leng was supported in part by NUS grant R-155-050-053-133. The authors thank the joint editors, the associate editor, and two referees for their careful reading and constructive comments. They also thank the associate editor, Bruce Brown, Chih-Ling Tsai, and Alec Stephenson for many detailed editorial suggestions, that led to a better presentation of the article.

Journal of the American Statistical Association
September 2007, Vol. 102, No. 479, Theory and Methods
DOI 10.1198/016214507000000509

$\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0d})^\top \in \mathbb{R}^d$ 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. 2007a,b):

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

The adaptive LASSO penalty used in (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 the 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. (2007a,b) proposed the foregoing 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 Least Squares Approximation

To motivate the 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. Later, we show 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

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

where $\dot{\mathcal{L}}_n(\cdot)$ and $\ddot{\mathcal{L}}_n(\cdot)$ are the first- and second-order derivatives of the loss function $\mathcal{L}_n(\cdot)$. Because $\tilde{\boldsymbol{\beta}}$ is the minimizer of $\mathcal{L}_n(\cdot)$, we know that $\dot{\mathcal{L}}_n(\tilde{\boldsymbol{\beta}}) = 0$. Thus the approximation (2) can be simplified to

$$n^{-1}\mathcal{L}_n(\boldsymbol{\beta}) \approx n^{-1}\mathcal{L}_n(\tilde{\boldsymbol{\beta}}) + \frac{1}{2}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^\top \left\{\frac{1}{n}\ddot{\mathcal{L}}_n(\tilde{\boldsymbol{\beta}})\right\}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}). \tag{3}$$

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

By ignoring the constant $\mathcal{L}_n(\tilde{\boldsymbol{\beta}})$ and the coefficient $1/2$, the objective function (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 $\mathrm{E}\{n^{-1}\ddot{\mathcal{L}}_n(\boldsymbol{\beta})\} \approx \boldsymbol{\Sigma}^{-1}$, 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 least squares function

$$(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) \tag{4}$$

as a simple approximation to the original loss $n^{-1}\mathcal{L}_n(\boldsymbol{\beta})$. We refer to (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 is still useful as long as $\hat{\boldsymbol{\Sigma}}$ is available.

2.3 The Least Squares Approximation Estimator

With LSA defined in (4), the original LASSO problem (1) can be rewritten as the asymptotically equivalent least squares problem

$$Q(\boldsymbol{\beta}) = (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) + \sum_{j=1}^{d} \lambda_j |\beta_j|, \tag{5}$$

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

The approximated LASSO objective function (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 (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), and so on. Together with the *lars* package in R, the proposed LSA method can be easily implemented. This further extends the practical applicability of the LASSO method and the LARS algorithm. Finally, because $\tilde{\boldsymbol{\beta}}$ used by LSA is not asymptotically efficient compared with the oracle, the resulting LSA estimator can be expected to be not efficient either. However, as we discuss in the next section, the LSA estimator is indeed as efficient as the 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 ($0 \le d_0 \le d$) predictors are relevant, that is, $\boldsymbol{\beta}_{0j} \neq 0$ for $j \le d_0$ and $\boldsymbol{\beta}_{0j} = 0$ for $j > d_0$. Let $\mathcal{S} = \{j_1, \ldots, j_{d^*}\}$ denote an arbitrary candidate model, that contains the j_1th, $\ldots$, j_{d^*}th ($1 \le d^* \le d$) predictors, and let $\mathcal{S}_F = \{1, \ldots, d\}$ denote the full model and $\mathcal{S}_T = \{1, \ldots, d_0\}$ denote the true model. 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 $\mathcal{S}$). Finally,

denote $\tilde{\boldsymbol{\beta}}_{\mathcal{S}} \in \mathbb{R}^d$ as the corresponding unpenalized estimator, that is,

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

This implies that $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}_{\mathcal{S}_F}$. On the other hand, by minimizing the objective function (4), another estimator,

$$\hat{\boldsymbol{\beta}}_{\mathcal{S}} = \underset{\{\boldsymbol{\beta}\in\mathbb{R}^d : \boldsymbol{\beta}_j=0, \forall j\notin\mathcal{S}\}}{\operatorname{argmin}} (\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}), \tag{6}$$

can be obtained. In general, $\tilde{\boldsymbol{\beta}}_{\mathcal{S}} \neq \hat{\boldsymbol{\beta}}_{\mathcal{S}}$. The estimator $\tilde{\boldsymbol{\beta}}_{\mathcal{S}}$ is obtained by minimizing the loss function $\mathcal{L}_n(\cdot)$, whereas $\hat{\boldsymbol{\beta}}_{\mathcal{S}}$ is produced by minimizing the LSA objective function (4). Although they are closely related, they are unlikely to be exactly the same. A similar difference also exists between $\tilde{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$.

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 $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}^{(\mathcal{S}_F)}$, $\boldsymbol{\beta}_0^{(\mathcal{S}_T)} \neq 0$, and the oracle estimator $\tilde{\boldsymbol{\beta}}_{\mathcal{S}_T}^{(\mathcal{S}_T)}$. In addition, assume that for any $\mathcal{S} \supset \mathcal{S}_T$,

$$\sqrt{n}\big(\tilde{\boldsymbol{\beta}}_{\mathcal{S}}^{(\mathcal{S})} - \boldsymbol{\beta}_0^{(\mathcal{S})}\big) \xrightarrow{d} \mathrm{N}(0, \boldsymbol{\Sigma}_{\mathcal{S}}) = \mathrm{N}(0, \boldsymbol{\Omega}_{\mathcal{S}}^{-1}) \tag{7}$$

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

$$\sqrt{n}\big(\tilde{\boldsymbol{\beta}}_{\mathcal{S}_T}^{(\mathcal{S}_T)} - \boldsymbol{\beta}_0^{(\mathcal{S}_T)}\big) \xrightarrow{d} \mathrm{N}\big(0, \boldsymbol{\Sigma}_{\mathcal{S}_T}\big) = \mathrm{N}\big(0, \boldsymbol{\Omega}_{\mathcal{S}_T}^{-1}\big). \tag{8}$$

Finally, for an artibrary $d \times d$ 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 that $\boldsymbol{\Omega}_{\mathcal{S}} = \boldsymbol{\Omega}_{\mathcal{S}_F}^{(\mathcal{S})}$.

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

Case 1. Consider the simplest situation, OLS regression with independent errors: $\boldsymbol{\Omega}_{\mathcal{S}} = \sigma^{-2}\operatorname{cov}(\mathbf{x}^{(\mathcal{S})})$, and σ^2 is the common variance of the error terms. Because $\operatorname{cov}(\mathbf{x}^{(\mathcal{S})})$ is a submatrix of $\operatorname{cov}(\mathbf{x}^{(\mathcal{S}_F)})$, 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 $\boldsymbol{\Omega}_{\mathcal{S}} = \tau\operatorname{cov}(\mathbf{x}^{(\mathcal{S})})$ 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:

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

For any overfitted model, the inverse asymptotic covariance matrix of the corresponding MLE (i.e., $\tilde{\boldsymbol{\beta}}_{\mathcal{S}}^{(\mathcal{S})}$) is $\boldsymbol{\Omega}_{\mathcal{S}} = \mathrm{E}\{p(\mathbf{x}^{(\mathcal{S}_T)})(1-p(\mathbf{x}^{(\mathcal{S}_T)}))\mathbf{x}^{(\mathcal{S})}\mathbf{x}^{(\mathcal{S})\top}\}$, a submatrix of $\boldsymbol{\Omega}_{\mathcal{S}_F} = \mathrm{E}\{p(\mathbf{x}^{(\mathcal{S}_T)})(1-p(\mathbf{x}^{(\mathcal{S}_T)}))\mathbf{x}^{(\mathcal{S}_F)}\mathbf{x}^{(\mathcal{S}_F)\top}\}$. Thus the covariance assumption is satisfied. Similarly, for any regression model with likelihood function 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 $\mathbf{x}$ through 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 $\boldsymbol{\Omega}_{\mathcal{S}} = 4f_e^2(0)\operatorname{cov}(\mathbf{x}^{(\mathcal{S})})$, where $f_e(0)$ is the probability density of the random noise e at the origin (Pollard 1991). Similar to Case 1, because $\operatorname{cov}(\mathbf{x}^{(\mathcal{S})})$ is the submatrix of $\operatorname{cov}(\mathbf{x}) = \operatorname{cov}(\mathbf{x}^{(\mathcal{S}_F)})$ 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 in which the covariance assumption is violated. For example, in Case 1, if the random noise depends on the predictor $\mathbf{x}$, then 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 specified otherwise.

3.2 The Oracle Property

We establish some relevant asymptotic theories in this section. Define $a_n = \max\{\lambda_j, j \le d_0\}$ and $b_n = \min\{\lambda_j, j > d_0\}$, and assume that the estimate $\hat{\boldsymbol{\Sigma}}$ is consistent, although its convergence rate can be arbitrary.

Theorem 1 ($\sqrt{n}$-consistency). If $\sqrt{n}a_n \xrightarrow{p} 0$, then $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} - \boldsymbol{\beta}_0 = O_p(n^{-1/2})$.

Here "$\xrightarrow{p}$" represents convergence in probability. Theorem 1 says that as long as the maximal regularization of the relevant predictors shrinks toward 0 faster than $n^{-1/2}$, the resulting LSA estimator is $\sqrt{n}$-consistent. The proof is given in the Appendix.

Next, we establish the consistency of the LSA estimator as a variable selection method. For convenience, define $\boldsymbol{\beta}_a = \boldsymbol{\beta}^{(\mathcal{S}_T)} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{d_0})^\top$ to be the vector corresponding to all of the nonzero coefficients and $\boldsymbol{\beta}_b = (\boldsymbol{\beta}_{d_0+1}, \ldots, \boldsymbol{\beta}_d)^\top$ to be the vector corresponding to all of the zero coefficients. Let $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},a}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},b}$ be their corresponding LSA estimators, that is, $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} = (\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},a}^\top, \hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},b}^\top)^\top$. Similarly, write $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_a^\top, \tilde{\boldsymbol{\beta}}_b^\top)^\top$ and $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0a}^\top, \boldsymbol{\beta}_{0b}^\top)^\top$. Then we have that $\boldsymbol{\beta}_{0a} \neq 0$, but $\boldsymbol{\beta}_{0b} = 0$.

Theorem 2 (Selection consistency). If $\sqrt{n}a_n \xrightarrow{p} 0$ and $\sqrt{n}b_n \xrightarrow{p} \infty$, then

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

Theorem 2 states that with probability tending to 1, all of the zero coefficients must be estimated as 0. Theorem 1 ensures the consistency of the estimators of the nonzero coefficients.

Together, Theorems 1 and 2 imply that the proposed LSA estimator can identify the true model consistently. The proof is given in the Appendix.

Theorem 3 (Oracle property). If $\sqrt{n}a_n \to 0$ and $\sqrt{n}b_n \to \infty$, then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},a} - \boldsymbol{\beta}_{0a}) \xrightarrow{d} \mathrm{N}\big(0, \boldsymbol{\Sigma}_{\mathcal{S}_T}\big).$$

Theorem 3 ensures the same asymptotic distribution of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},a}$ as that of the oracle. Consequently, $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},a}$ can be as efficient as the oracle. The proof is given in the 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, even 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 appropriate specification of the tuning parameters. For practical implementation, the problem of simultaneously tuning so many regularization parameters (i.e., λ_j) is challenging. Intuitively, some model selection criteria, such as generalized cross-validation (GCV), the Akaike information criterion, or the Bayes information criterion (BIC) can be used to choose $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)^\top$ through an exhaustive grid search. But 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):

$$\lambda_j = \lambda_0 |\tilde{\boldsymbol{\beta}}_j|^{-\gamma}, \tag{9}$$

where $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1, \ldots, \tilde{\boldsymbol{\beta}}_d)^\top$ is the unpenalized original estimator and $\gamma > 0$ is some prespecified 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. Because $\tilde{\boldsymbol{\beta}}_j$ is $\sqrt{n}$-consistent, we can verify that the tuning parameter defined in (9) satisfies all of the technical conditions needed by Theorems 1, 2, and 3, as long as $n^{1/2}\lambda_0 \to 0$ and $n^{(1+\gamma)/2}\lambda_0 \to \infty$. Thus it suffices to select $\lambda_0 \in \mathbb{R}^+ = [0, \infty)$ only.

3.3 The Bayes Information Criterion

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

$$\mathrm{BIC}_{\boldsymbol{\lambda}} = (\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\beta}})^\top \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\beta}}) + \log n \times df_{\boldsymbol{\lambda}}/n, \tag{10}$$

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, Hastie, and Tibshirani 2004).

The estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ naturally defines a candidate model $\mathcal{S}_{\boldsymbol{\lambda}} = \{j : \hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},j} \neq 0\}$. Thus 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 $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)^\top$, whereas $\hat{\boldsymbol{\beta}}_{\mathcal{S}_{\boldsymbol{\lambda}}}$ is the unpenalized estimator determined by (6) 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 (6), the least squares loss produced by $\hat{\boldsymbol{\beta}}_{\mathcal{S}_{\boldsymbol{\lambda}}}$ must be no larger than that of the shrinkage estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$, that is,

$$(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\beta}})^\top \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\beta}}) \geq \big(\hat{\boldsymbol{\beta}}_{\mathcal{S}_{\boldsymbol{\lambda}}} - \tilde{\boldsymbol{\beta}}\big)^\top \hat{\boldsymbol{\Sigma}}^{-1} \big(\hat{\boldsymbol{\beta}}_{\mathcal{S}_{\boldsymbol{\lambda}}} - \tilde{\boldsymbol{\beta}}\big). \tag{11}$$

According to whether the resulting model $\mathcal{S}_{\boldsymbol{\lambda}}$ is underfitted, correctly fitted, or overfitted, we can partition $\mathbb{R}^{d+}$ into the following three mutually exclusive regions:

$$\mathbb{R}^d_- = \{\boldsymbol{\lambda} \in \mathbb{R}^{d+} : \mathcal{S}_{\boldsymbol{\lambda}} \not\supset \mathcal{S}_T\},$$

$$\mathbb{R}^d_0 = \{\boldsymbol{\lambda} \in \mathbb{R}^{d+} : \mathcal{S}_{\boldsymbol{\lambda}} = \mathcal{S}_T\},$$

and

$$\mathbb{R}^d_+ = \{\boldsymbol{\lambda} \in \mathbb{R}^{d+} : \mathcal{S}_{\boldsymbol{\lambda}} \supset \mathcal{S}_T, \mathcal{S}_{\boldsymbol{\lambda}} \neq \mathcal{S}_T\}.$$

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 $\log n/\sqrt{n}$. Because $\boldsymbol{\lambda}_n$ satisfies the conditions in Theorems 1 and 2, it follows that $\mathcal{S}_{\boldsymbol{\lambda}_n} = \mathcal{S}_T$ with probability tending to 1. 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 (6) has an explicit solution, and $\hat{\boldsymbol{\beta}}_{\mathcal{S}} \xrightarrow{p} \boldsymbol{\beta}_{\mathcal{S}}$ for some $\boldsymbol{\beta}_{\mathcal{S}}$. Similarly, we can verify that $\boldsymbol{\beta}_{\mathcal{S}} = \boldsymbol{\beta}_0$ for any overfitted model $\mathcal{S} \supset \mathcal{S}_T$. However, the situation is different for an underfitted model $\mathcal{S} \not\supset \mathcal{S}_T$. Because we are forcing some nonzero coefficients to be 0, we must have that $\boldsymbol{\beta}_{\mathcal{S}} \neq \boldsymbol{\beta}_0$. Then the consistency of the proposed BIC criterion (10) can be established.

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

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. The proof is given in the Appendix.

Remark 2. It can be shown that Theorem 4 holds independent of the covariance assumption. Furthermore, if the factor $\log n$ in (10) is replaced by some other factor γ_n, Theorem 4 still holds as long as $\gamma_n/n \to 0$ and $\gamma_n \to \infty$.

4. NUMERICAL STUDIES

We present extensive numerical studies to demonstrate LSA's finite-sample performance. All studies were conducted in R with the *lars* package (Efron et al. 2004). This package (and many others used here) can be downloaded at *http://cran.r-project.org/*. The LSA method uses $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}$ as 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 (10) for LSA (unless specified otherwise) 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).

Table 1. Simulation results for logistic regression

Sample size	Estimation method	RME		MS		CM	
		Median	(SE)	Mean	(SE)	Mean	(SE)
200	LSA	.504	(.015)	3.178	(.026)	.798	(.018)
	PH	.544	(.017)	3.272	(.033)	.716	(.020)
400	LSA	.417	(.013)	3.130	(.018)	.888	(.014)
	PH	.566	(.024)	3.092	(.023)	.846	(.016)

Table 2. Simulation results for Cox's model

Sample size	Estimation method	RME		MS		CM	
		Median	(SE)	Mean	(SE)	Mean	(SE)
100	LSA	.440	(.038)	3.394	(.031)	.644	(.021)
	PH	.451	(.047)	3.342	(.033)	.634	(.022)
200	LSA	.575	(.065)	3.208	(.022)	.822	(.017)
	PH	.643	(.096)	3.182	(.022)	.824	(.017)

Example 1 (Logistic regression). In this example, independent observations with binary response are generated according to the 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 $.5^{|j_1 - j_2|}$. 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 a path is identified by its corresponding BIC. For Park and Hastie's algorithm, the BIC is defined to be $-2\log L + df_{\boldsymbol{\lambda}} \times \log(n)$, where L stands for the log-likelihood of the data. The final estimate is designated the PH estimator. Because 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.

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

$$h(t_i|\mathbf{x}_i) = \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_0), \tag{12}$$

where t_i is the survival time from the ith subject and $\boldsymbol{\beta}_0 = (.8, 0, 0, 1, 0, 0, .6, 0)^\top$. According to (12), the survival time of the ith subject is distributed exponentially with mean $\exp(-\mathbf{x}_i^\top \boldsymbol{\beta}_0)$. Again, the covariate x_{ij} is generated in the same manner as in 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 distributed uniformly 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 easily implemented. As before, we compare the performance of the LSA estimator with that of the PH estimator, using the BIC $-2\log L + df_{\boldsymbol{\lambda}} \log(n)$ for the latter, where L is the partial likelihood (Cox 1972). The simulation results, summarized in Table 2, confirm the similar performance of the LSA and PH estimators.

Example 3 (LAD regression). As our third example, we revisit the LAD–LASSO problem studied by Wang et al. (2007a,b). Data are simulated according to

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

where $\sigma = 1$, $\mathbf{x}_i$ 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% coming from a standard Cauchy distribution. Because Wang et al. (2007a) 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 can be easily implemented. The BIC used by the WLJ estimator was given by Hurvich and Tsai (1990, eq. 8). The results, summarized in Table 3, indicate that the LSA estimator has slightly better performance than 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 $\tilde{\boldsymbol{\beta}}$ is obtained by minimizing the 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, sec. 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.

Example 5 (Buckley–James estimator). As our fifth simulation example, we consider the Buckley–James estimator (Buckley and James 1979), an attractive alternative to the Cox proportional hazards model. Survival data are generated according to the 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 ith subject, $\boldsymbol{\beta}_0 = (.8, -.5, 0, 0, .5, 0, 0, 0)^\top$, $\sigma = .5$, and the covariate $\mathbf{x}_i$ is gen-

Table 3. Simulation results for LAD regression

Sample size	Estimation method	RME		MS		CM	
		Median	(SE)	Mean	(SE)	Mean	(SE)
100	LSA	.463	(.017)	3.022	(.007)	.980	(.006)
	WLJ	.516	(.018)	3.262	(.028)	.790	(.018)
200	LSA	.429	(.013)	3.010	(.004)	.990	(.004)
	WLJ	.453	(.015)	3.148	(.021)	.876	(.015)

Table 4. Simulation results for Tukey's biweight regression

Sample size	RME		MS		CM	
	Median	(SE)	Mean	(SE)	Mean	(SE)
50	.528	(.015)	3.342	(.030)	.750	(.019)
100	.495	(.015)	3.160	(.020)	.864	(.015)
200	.455	(.015)	3.076	(.013)	.928	(.012)

erated 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 easily implemented. The asymptotic covariance matrix of the Buckley–James estimator is of the form $\mathbf{\Sigma} = \mathbf{W}^{-1}\mathbf{V}\mathbf{W}^{-1}$ 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.

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 $\hat{\mathbf{\Sigma}}^{-1}$. An asymptotically valid 95% confidence interval is then constructed through the normal approximation. The actual coverage probabilities (in percentages) are reported in Table 6, along with those of the adaptive LASSO (aLASSO) and oracle (ORACLE).

For relatively large sample sizes (e.g., $n = 200$), the performance of the LSA is comparable to that of the oracle, further confirming our asymptotic theory. With smaller sample sizes (e.g., $n = 100$), the LSA's performance slightly worsens but is still practically acceptable. The diminished 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 were also conducted for Examples 1, 3, and 4, with similar findings.

Example 7 (South Africa Heart Disease Data). As our first real data example, we consider the South Africa Heart Disease Data of 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, 9 predictors are associated with a binary response variable, which indicates the presence or absence of myocardial infarction for 462 subjects. To compare the differences in the original LASSO, adaptive LASSO, and proposed LSA estimators, we plot their solution paths on the left side of Figure 1. For the PH estimates, we set the arc length at .1 (Park and Hastie 2006a). Little difference can be seen between the adaptive PH estimate and our LSA estimate. Both give a model with the same set of nonzero coefficients. In 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 were collected by the Mayo Clinic between January 1974 and May 1984. The data include 312 randomized patients with 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; the results are presented on the right side of Figure 1. Once again, little difference can be seen 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 an anonymous referee, it would be more attractive if the proposed LSA method could be potentially useful in the situation where $d \gg n$. Consequently, we devote this example to evaluate this possibility using the leukemia microarray data set (Golub et al. 1999), which comprises 72 samples. 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 3,571 genes (i.e., $d = 3{,}571$) (see Dudoit, Fridlyand, and Speed 2002). Following Golub et al. (1999), the first 38 (i.e., $n = 38$) samples are used for training, and the remaining 34 samples are reserved for testing. Clearly, the predictor dimension ($d = 3{,}571$) is substantially larger than the sample size ($n = 38$). Thus 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 $(\tilde{\boldsymbol{\beta}}, \hat{\mathbf{\Sigma}})$ by their corresponding regularized estimates $(\tilde{\boldsymbol{\beta}}_\kappa, \hat{\mathbf{\Sigma}}_\kappa)$, where $\kappa > 0$ is the tuning parameter used by the L_2 penalty. To avoid the double-shrinkage effect (Zou and Hastie 2006), 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 tenfold cross-validation. Our experience suggests that

Table 5. Simulation results for the Buckley–James estimator

Sample size	RME		MS		CM	
	Median	(SE)	Mean	(SE)	Mean	(SE)
50	.697	(.021)	3.326	(.030)	.712	(.020)
100	.582	(.020)	3.166	(.020)	.860	(.016)
200	.553	(.014)	3.108	(.015)	.900	(.013)

Table 6. Confidence interval coverage probability for Cox's model

Estimation method	Sample size					
	$n = 100$			$n = 200$		
	$\boldsymbol{\beta}_{01}$	$\boldsymbol{\beta}_{04}$	$\boldsymbol{\beta}_{07}$	$\boldsymbol{\beta}_{01}$	$\boldsymbol{\beta}_{04}$	$\boldsymbol{\beta}_{07}$
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

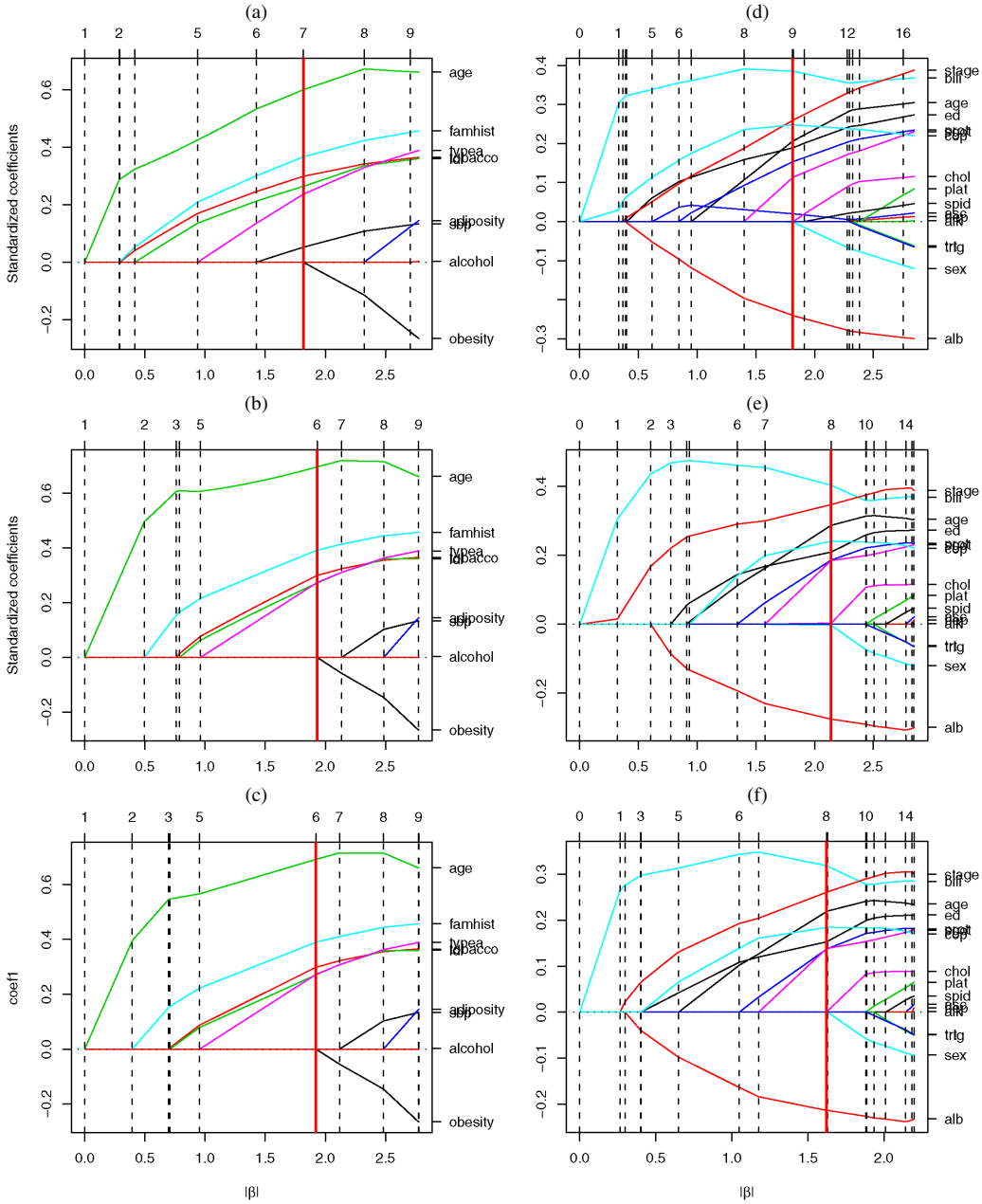

Figure 1. South Africa Heart Disease Data (left side) and Primary Biliary Cirrhosis Data (right side). (a) and (d) The solution paths for the original LASSO by the PH algorithm. (b) and (e) The solution paths for the adaptive LASSO by the PH algorithm. (c) and (f) The solution paths by the LSA. The vertical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

the BIC criterion (10) 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 only as a very preliminary evaluation regarding the potential usefulness of the LSA method in the situation where $d \gg n$. Theoretically, we are not clear whether the covariance assumption is satisfied here. Computationally, we find that the small-sample performance of the tenfold cross-validation method could be unstable. Despite our encouraging findings, however, 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 an anonymous referee, for a given data set with fixed sample size n and unknown parameter $\boldsymbol{\beta}_0$, it is difficult to determine whether the sample size is sufficient. Thus it is of interest to evaluate the finite-sample performance of the LSA method under some local perturbation. To do this, we follow the idea of Leeb and Pötscher (2006) and replicate Example 1 but 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 re-

Table 7. Simulation results for logistic regression under local perturbation

Sample size	Estimation method	Coverage probability of 95% confidence interval (%) β_1	β_4	β_7	Percentage of correct models
200	LSA	9.2	3.8	4.2	1.0
	aLASSO	9.0	3.8	4.2	.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

sults show that both LSA and adaptive LASSO perform poorly. These results further confirm the theoretical findings of Leeb and Pötscher (2005) stating 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.

Second, the covariance assumption can be violated (see, e.g., Example 5 in Sec. 5). However, even if the covariance assumption is violated, the LSA method remains 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: PROOFS

Proof of Theorem 1

Note that he LSA objective function $Q(\boldsymbol{\beta})$ given in (5) is a strictly convex function in $\boldsymbol{\beta}$. Thus as long as we can show that (5) has a $\sqrt{n}$-consistent local minimizer, it must be $\sqrt{n}$-consistent global minimizer. Thus the theorem's conclusion follows immediately. Following 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

$$\liminf_n P\Big\{\inf_{\mathbf{u}\in\mathbb{R}^d:\,\|\mathbf{u}\|=C} Q\big(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}\big) > Q(\boldsymbol{\beta}_0)\Big\} > 1-\epsilon. \tag{A.1}$$

Let $\mathbf{u}= (u_1,\ldots,u_d)^\top$. Then, by the definition of (5), simple algebra shows that

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

where the second equality holds because $\boldsymbol{\beta}_{0j}=0$ for any $j > d_0$ by the model assumption. In addition, according to the theorem's condition $\sqrt{n}a_n \xrightarrow{p} 0$, 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 $\tau_{\min}(\hat{\boldsymbol{\Sigma}}^{-1})C^2 \xrightarrow{p} \tau_{\min}(\boldsymbol{\Sigma}^{-1})C^2$, where $\tau_{\min}(M)$ refers to the minimal eigenvalue of M. It follows then, with probability tending to 1, that the first term in (A.2) is uniformly larger than $.5\tau_{\min}(\boldsymbol{\Sigma}^{-1})C^2$, which is quadratic in C. On the other hand, the second term in (A.2) is uniformly bounded by $C\|\hat{\boldsymbol{\Sigma}}^{-1}\sqrt{n}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})\|$, which is linear in C with coefficient $\|\hat{\boldsymbol{\Sigma}}^{-1}\sqrt{n}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})\| = O_p(1)$. Therefore, as long as the constant C is sufficiently large, the first term will always dominate the other two terms with arbitrarily large probability. This implies the inequality (A.1), and the proof is completed.

Proof of Theorem 2

We need only show that $P(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},j}=0)\to 1$ for any $d_0 < j \le d$. If $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},j}\neq 0$ for some $d_0 < j \le d$, then

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

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

Proof of Theorem 3

For convenience purposes, first decompose the asymptotic covariance matrix $\boldsymbol{\Sigma}$ into the following block matrix form:

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\mathcal{S}_F} = \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix},$$

where $\boldsymbol{\Sigma}_{11}$ is the first $d_0\times d_0$ submatrix. It then follows that $\boldsymbol{\Sigma}_{11} = \boldsymbol{\Sigma}^{(\mathcal{S}_T)}$. Similarly, its inverse matrix $\boldsymbol{\Omega} = \boldsymbol{\Omega}_{\mathcal{S}_F} = \boldsymbol{\Sigma}^{-1}$ can also be partitioned as

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

The estimators $(\hat{\boldsymbol{\Sigma}}_{ij}, \hat{\boldsymbol{\Omega}}_{ij})$ $(i,j\in\{1,2\})$ can be defined similarly. According to Theorem 2, with probability tending to 1,

$\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},b} = 0$; thus $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},a}$ must be the global minimizer of the objective function

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

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 $1 \le j \le d_0$. Thus the following normal equation must be satisfied:

$$\begin{aligned} 0 &= \frac{1}{2} \times \frac{\partial Q_0(\boldsymbol{\beta}_a)}{\partial \boldsymbol{\beta}_a}\bigg|_{\boldsymbol{\beta}_a = \hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},a}} \\ &= \hat{\boldsymbol{\Omega}}_{11}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},a} - \tilde{\boldsymbol{\beta}}_a) - \hat{\boldsymbol{\Omega}}_{12}\tilde{\boldsymbol{\beta}}_b + D(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},a}), \end{aligned} \tag{A.4}$$

where $D(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},a})$ is a d_0-dimensional vector with its jth component given by $.5\lambda_j \operatorname{sgn}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},j})$. Note that (A.4) implies that

$$\begin{aligned} &\sqrt{n}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},a} - \boldsymbol{\beta}_{0a}) \\ &\quad = \sqrt{n}(\tilde{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_{0a}) + \hat{\boldsymbol{\Omega}}_{11}^{-1}\hat{\boldsymbol{\Omega}}_{12}(\sqrt{n}\tilde{\boldsymbol{\beta}}_b) - \hat{\boldsymbol{\Omega}}_{11}^{-1} \times \sqrt{n} D(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},a}) \\ &\quad = \sqrt{n}(\tilde{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_{0a}) + \hat{\boldsymbol{\Omega}}_{11}^{-1}\hat{\boldsymbol{\Omega}}_{12}(\sqrt{n}\tilde{\boldsymbol{\beta}}_b) + o_p(1) \qquad \text{(A.5)} \\ &\quad = \sqrt{n}(\tilde{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_{0a}) + \boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}(\sqrt{n}\tilde{\boldsymbol{\beta}}_b) + o_p(1), \qquad \text{(A.6)} \end{aligned}$$

where the equality (A.5) is due mainly to the fact that $\sqrt{n} \times D(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},a}) = o_p(1)$ because $\sqrt{n}\lambda_j < \sqrt{n}a_n \xrightarrow{p} 0$ for any $j \le d_0$, 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

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

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

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

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

$$\boldsymbol{\Sigma}_{(11)}^{-1} = (\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})^{-1} = \boldsymbol{\Omega}_{11}.$$

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

Proof of Theorem 4

Following a similar idea of Wang et al. (2007b), for any $\mathcal{S}_{\boldsymbol{\lambda}} \neq \mathcal{S}_T$, we consider two different cases according to whether the model is underfitted or overfitted. We show that the theorem conclusion is valid in both cases.

Case A.1 (Underfitted model). First, note that $\boldsymbol{\lambda}_n$ satisfies the regularity conditions specified by Theorem 1. Consequently, the resulting estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_n}$ is $\sqrt{n}$-consistent, which immediately implies that its associated BIC value is of the order $o_p(1)$. In contrast, due to the fact that $\boldsymbol{\beta}_{\mathcal{S}} \neq \boldsymbol{\beta}_0$ for any $\mathcal{S} \not\supset \mathcal{S}_T$, we know that

$$\begin{aligned} \mathrm{BIC}_{\boldsymbol{\lambda}} &= (\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\beta}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\beta}}) + df_{\boldsymbol{\lambda}} \times \log(n)/n \\ &\ge (\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\beta}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\beta}}) \\ &\ge (\hat{\boldsymbol{\beta}}_{\mathcal{S}_{\boldsymbol{\lambda}}} - \tilde{\boldsymbol{\beta}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\beta}}_{\mathcal{S}_{\boldsymbol{\lambda}}} - \tilde{\boldsymbol{\beta}}) \qquad \text{(A.7)} \\ &\ge \min_{\mathcal{S} \not\supset \mathcal{S}_T} (\hat{\boldsymbol{\beta}}_{\mathcal{S}} - \tilde{\boldsymbol{\beta}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\beta}}_{\mathcal{S}} - \tilde{\boldsymbol{\beta}}) \\ &\xrightarrow{p} \min_{\mathcal{S} \not\supset \mathcal{S}_T} (\boldsymbol{\beta}_{\mathcal{S}} - \boldsymbol{\beta})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}_{\mathcal{S}} - \boldsymbol{\beta}) > 0, \qquad \text{(A.8)} \end{aligned}$$

where the inequality (A.7) is due to (11) and the last inequality in (A.8) is due to the fact that $\boldsymbol{\Sigma}$ is positive definite and $\boldsymbol{\beta}_{\mathcal{S}} \neq \boldsymbol{\beta}_0$ for any $\mathcal{S} \not\supset \mathcal{S}_T$. Consequently, with probability tending to 1, we must have that $\inf_{\boldsymbol{\lambda} \in \mathbb{R}_-^d} \mathrm{BIC}_{\boldsymbol{\lambda}} > \mathrm{BIC}_{\boldsymbol{\lambda}_n}$. Thus it follows immediately that $P(\inf_{\boldsymbol{\lambda} \in \mathbb{R}_-^d} \mathrm{BIC}_{\boldsymbol{\lambda}} > \mathrm{BIC}_{\boldsymbol{\lambda}_n}) \to 1$.

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

$$\begin{aligned} &n\big(\mathrm{BIC}_{\boldsymbol{\lambda}} - \mathrm{BIC}_{\boldsymbol{\lambda}_n}\big) \\ &\quad = n(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\beta}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\beta}}) \\ &\qquad - n\big(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_n} - \tilde{\boldsymbol{\beta}}\big)^\top \hat{\boldsymbol{\Sigma}}^{-1}\big(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_n} - \tilde{\boldsymbol{\beta}}\big) + (df_{\boldsymbol{\lambda}} - d_0)\log n \\ &\quad \ge n\big(\hat{\boldsymbol{\beta}}_{\mathcal{S}_{\boldsymbol{\lambda}}} - \tilde{\boldsymbol{\beta}}\big)^\top \hat{\boldsymbol{\Sigma}}^{-1}\big(\hat{\boldsymbol{\beta}}_{\mathcal{S}_{\boldsymbol{\lambda}}} - \tilde{\boldsymbol{\beta}}\big) \\ &\qquad - n\big(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_n} - \tilde{\boldsymbol{\beta}}\big)^\top \hat{\boldsymbol{\Sigma}}^{-1}\big(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_n} - \tilde{\boldsymbol{\beta}}\big) + (df_{\boldsymbol{\lambda}} - d_0)\log n \\ &\quad \ge n\big(\hat{\boldsymbol{\beta}}_{\mathcal{S}_{\boldsymbol{\lambda}}} - \tilde{\boldsymbol{\beta}}\big)^\top \hat{\boldsymbol{\Sigma}}^{-1}\big(\hat{\boldsymbol{\beta}}_{\mathcal{S}_{\boldsymbol{\lambda}}} - \tilde{\boldsymbol{\beta}}\big) \\ &\qquad - n\big(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_n} - \tilde{\boldsymbol{\beta}}\big)^\top \hat{\boldsymbol{\Sigma}}^{-1}\big(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_n} - \tilde{\boldsymbol{\beta}}\big) + \log n \\ &\quad \ge \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}}) \\ &\qquad - n\big(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_n} - \tilde{\boldsymbol{\beta}}\big)^\top \hat{\boldsymbol{\Sigma}}^{-1}\big(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_n} - \tilde{\boldsymbol{\beta}}\big) + \log n, \qquad \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 (7). 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. Finally, note that the last term in (A.9) diverges to $+\infty$ as $n \to \infty$. As a result, we must have $P(\inf_{\boldsymbol{\lambda} \in \mathbb{R}_+^d} \mathrm{BIC}_{\boldsymbol{\lambda}} > \mathrm{BIC}_{\boldsymbol{\lambda}_n}) \to 1$. This completes Case A.2, and thus the proof is completed.

[Received May 2006. Revised March 2007.]

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*, Ser. 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, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999), "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, 286, 531–537.

Huber, P. (1981), *Robust Estimation*, New York: Wiley.

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, Yale University, Dept. of Statistics.

——— (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, 1273–1284.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, New York: Chapman & Hall.

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, 389–404.

Park, M. Y., and Hastie, T. (2006a), "An L1 Regularization-Path Algorithm for Generalized Linear Models," manuscript, Stanford University, Dept. of Statistics.

——— (2006b), "Penalized Logistic Regression for Detecting Gene Interactions," manuscript, Stanford University, Dept. of Statistics.

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*, Ser. B, 58, 267–288.

——— (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. (2007a), "Robust Regression Shrinkage and Consistent Variable Selection via the LAD–LASSO," *Journal of Business & Economic Statistics*, to appear.

Wang, H., Li, R., and Tsai, C. L. (2007b), "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*, Ser. B, 69, 143–161.

Zhao, P., and Yu, B. (2004), "Boosted LASSO," Technical Report 678, University of California Berkeley, Dept. of Statistics.

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*, Ser. B, 67, 301–320.

Zou, H., Hastie, T., and Tibshirani, R. (2004), "On the 'Degrees of Freedom' of LASSO," technical report, Stanford University, Dept. of Statistics.