Statistica Sinica **19** (2009), 449-471

BOOTSTRAP-BASED PENALTY CHOICE FOR THE LASSO, ACHIEVING ORACLE PERFORMANCE

Peter Hall[1], Eun Ryung Lee[2] and Byeong U. Park[2]

[1] *The University of Melbourne and* [2] *Seoul National University*

Abstract: In theory, if penalty parameters are chosen appropriately then the lasso can eliminate unnecessary variables in prediction problems, and improve the performance of predictors based on the variables that remain. However, standard methods for tuning-parameter choice, for example techniques based on the bootstrap or cross-validation, are not sufficiently accurate to achieve this level of precision. Until Zou's (2006) proposal for an inversely-weighted lasso, this difficulty led to speculation that it might not be possible to achieve oracle performance using the lasso. In the present paper we show that a straightforward application of the m-out-of-n bootstrap produces adaptive penalty estimates that confer oracle properties on the lasso. The application is of interest in its own right since, unlike many uses of the m-out-of-n bootstrap, it is not designed to estimate a non-normal distribution; the limiting distributions of regression parameter estimators are normal. Instead, the m-out-of-n bootstrap overcomes the tendency of the standard bootstrap to confound the errors committed in determining whether or not a parameter value is zero, with estimation errors for nonzero parameters.

Key words and phrases: Adaptive inference, bootstrap, m-out-of-n bootstrap, optimality properties, prediction, regression, variable selection.

1. Introduction

Tibshirani's (1996) lasso has proved particularly popular for both variable selection and parameter estimation. In the first of these settings it eliminates, or at least downweights, explanatory variables that it assesses to be of only minor influence, and in the second it offers the advantage, over ordinary least squares, of not degrading performance by estimation of redundant parameters. However, achieving these outcomes requires careful choice of the smoothing parameter, generally a vector of nonnegative penalties.

Standard cross-validation or bootstrap methods fail to achieve oracle performance in this problem. In particular, they do not produce parameter estimators that are asymptotically negligible, i.e. that are $o_p(n^{-1/2})$, where n denotes sample size, when true parameter values are zero. The reason is that conventional algorithms are confused by small errors, of order $n^{-1/2}$, that are implicit in parameter estimates even when the true values vanish. Unless the oracle tells us which parameters are zero, so that we have information from outside the sample and can drop those components from the model, we do not achieve (using conventional empirical methods) the good properties for which the lasso is known.

The difficulties arise because the level at which standard methods commit errors in determining whether a parameter value is zero, is the same as the level of accuracy at which we are conducting inference; the size is $n^{-1/2}$ in both cases. However, if we employ the m-out-of-n bootstrap, where m is of strictly smaller order of magnitude than n, then the size of the errors we commit in determining whether a parameter value is zero remains at $n^{-1/2}$, but the level of accuracy with which we are estimating the parameters is strictly larger, in fact $m^{-1/2}$. Therefore, as we shall detail in this paper, the former error does not confound the latter, and oracle performance can be achieved.

We shall demonstrate that, using the m-out-of-n bootstrap to choose penalties adaptively, the number of zero parameter values, and their locations in the parameter vector, are estimated consistently. The estimators of nonzero parameter values enjoy the asymptotic variance they would have if the zero values were eliminated from the model. We also develop an adaptive algorithm for choosing empirically the value of m.

Our method can be compared with alternative approaches discussed by, for example, Fan and Li (2001), who introduced a “smoothly clipped standard deviation” (SCAD) penalty and shed doubt on the potential for achieving oracle properties in the case of the lasso; Meinshausen and Bühlmann (2006), who pointed to the conflicts inherent in optimal estimation and accurate variable selection; and Zou (2006), who gave necessary conditions for a singly-penalized version of the lasso to be consistent. Zou also discussed the adaptive form of the lasso, where there is a different penalty for each parameter, and introduced a method which achieves oracle properties in that setting. Zou’s (2006) approach was based on weights that are inversely proportional to powers of true parameter values, with a single penalty parameter. By way of contrast, our method eliminates the weights and, in a p-variate problem, chooses all p penalties together.

There is a very large literature on, or closely related to, the lasso. It includes other contributions to L_1 penalty methods, for example in connection with basis pursuit (e.g., Chen, Donoho and Saunders (2001) and Ferris, Voelker and Zhang (2004)); work on the nonnegative garotte (e.g., Breiman (1995) and Gao (1998)); research on soft thresholding (e.g., Donoho, Johnstone, Kerkyacharian and Picard (1995)); and contributions to inference under sparsity (e.g., Donoho and Huo (2001), Donoho and Elad (2003), Tropp (2005) and Donoho (2006a,b)).

2. Model and Bootstrap Methods

2.1. Model and estimator

Assume that data pairs (x_i, Y_i), for $1 \le i \le n$, are generated under the model

$$Y_i = \beta_0 + x_i^{\mathrm{T}} \beta + \epsilon_i \,, \tag{2.1}$$

where x_i and $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ are p-vectors; Y_i, β_0 and ϵ_i are scalars; and, conditional on $x_1, \ldots, x_n$, the experimental errors ϵ_i are independent and identically distributed with zero mean and variance σ^2.

We suppose too that the covariates are centered at their empirical means, so that

$$\sum_{i=1}^{n} x_i = 0\,. \tag{2.2}$$

This condition is imposed without loss of generality, since inference is conducted conditionally on the covariates, and, in particular, the assumption that the x_i's are independent random vectors is not required. Note that recentering to achieve (2.2) requires $x_i = x_{ni}$ to depend on n, although we usually suppress the subscript n.

Under (2.2) we can, and shall, estimate β_0 as $\bar{Y} = n^{-1} \sum_i Y_i$. To estimate the true values $\beta_1^0, \ldots, \beta_p^0$ of $\beta_1, \ldots, \beta_p$, let $\lambda = (\lambda_1, \ldots, \lambda_p)^{\mathrm{T}}$ denote a vector of nonnegative components, and define $\hat{\beta} = \hat{\beta}_\lambda = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^{\mathrm{T}}$ to be the minimizer of

$$S_\lambda(\beta) = \sum_{i=1}^{n} \left(Y_i - \bar{Y} - x_i^{\mathrm{T}}\beta\right)^2 + \sum_{j=1}^{p} \lambda_j\, |\beta_j|\,. \tag{2.3}$$

Let $\mathcal{R}$ be the set of integers j for which $\beta_j^0 = 0$, and write $n^{-1}\,\Sigma^1$ for the limiting covariance matrix of the least-squares estimator after the model (2.1) has been reduced by eliminating β_j for all $j \in \mathcal{R}$. Denote by $\widehat{\mathcal{R}}$ the set of j such that $\hat{\beta}_j = 0$. The estimator $\hat{\beta}$ is said to have the oracle property if,

$n^{1/2}\,(\hat{\beta} - \beta^0)$, after components corresponding to elements of $\mathcal{R}$ have been removed, is asymptotically normal $\mathrm{N}(0, \Sigma^1)$, and $P(\widehat{\mathcal{R}} = \mathcal{R}) \to 1$. (2.4)

2.2. Bootstrap algorithm

Let $\tilde{\beta}$ denote a root-n consistent "pilot estimator" of β^0. The simplest choice of $\tilde{\beta}$ is the standard least-squares estimator, equal to the minimizer of $S_\lambda(\beta)$ when λ is a vector of zeros, and we use that estimator in Section 4. Compute the residuals, $\hat{\epsilon}_i = Y_i - \bar{Y} - x_i^{\mathrm{T}}\tilde{\beta}$; put $\tilde{\epsilon}_i = \hat{\epsilon}_i - n^{-1}\sum_j \hat{\epsilon}_j$ (if $\tilde{\beta}$ is the standard least-squares estimator then $\tilde{\epsilon}_i = \hat{\epsilon}_i$); given $m \in \{1, \ldots, n\}$, obtain $\epsilon_1^*, \ldots, \epsilon_m^*$ by sampling randomly, with replacement, from $\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_n$ (recall that we are using the m-out-of-n bootstrap); and define $Y_i^* = \bar{Y} + x_i^{\mathrm{T}}\tilde{\beta} + \epsilon_i^*$ for $1 \le i \le m$. Here, $x_1, \ldots, x_n$ are unchanged from their appearance in the dataset $(x_1, Y_1), \ldots, (x_n, Y_n)$ introduced in Section 2.1. Writing $\lambda = (\lambda_1, \ldots, \lambda_p)^{\mathrm{T}}$ for a vector of nonnegative numbers, let $\beta = \hat{\beta}^*$ be the minimizer of

$$S_\lambda^*(\beta) = \sum_{i=1}^{m} \left(Y_i^* - \bar{Y}^* - x_i^{\mathrm{T}}\beta\right)^2 + \sum_{j=1}^{p} \lambda_j\, |\beta_j|\,.$$

Let $\mathcal{Z} = \{(x_1, Y_1), \ldots, (x_n, Y_n)\}$ denote the dataset, and define

$$\widehat{\mathrm{MSE}}(\lambda) = E\big(\|\hat\beta^* - \tilde\beta\|^2 \,\big|\, \mathcal{Z}\big)\,,$$

an estimator of $\mathrm{MSE}(\lambda) = E(\|\hat\beta - \beta\|^2)$. Put

$$\tilde\lambda = \mathrm{argmin}\, \widehat{\mathrm{MSE}}(\lambda)\,, \quad \hat\lambda = \Big(\frac{n}{m}\Big)^{\frac{1}{2}}\, \tilde\lambda\,, \tag{2.5}$$

where of course the minimum is taken over p-vectors λ with nonnegative components.

The correctness of the normalization $(n/m)^{1/2}$ in (2.5) derives from the fact that this standardization is appropriate in asymptotic results under local perturbation models; see Section 3.1, and particularly Theorem 1. The relevance of this viewpoint can be appreciated when it is noted that $\hat\beta_j$ is, in effect, a local perturbation of the true value, β_j^0, of β_j, on a scale of $n^{-1/2}$.

Our final estimator of β is the value $\bar\beta = (\bar\beta_1, \ldots, \bar\beta_p)^{\mathrm{T}}$ that minimizes $S_{\hat\lambda}(\beta)$. We show in Section 3.3 that $\bar\beta$ is first-order optimal. In particular, just as in the case of the estimator $\hat\beta_{\mathrm{opt}}$ which uses the theoretically optimal, but unknown, value of λ, the jth component of $\bar\beta$ converges to zero faster than $n^{-1/2}$ if $\beta_j^0 = 0$, and the components of $\bar\beta_j$ that correspond to nonzero β_j^0's have the same first-order asymptotic properties as the least-squares estimators constrained to those components.

A bootstrap approach alternative to that given two paragraphs above would be to resample the design variables x_i, as well as the centered residuals $\tilde\epsilon_i$. This would be appropriate if the x_i's were being treated as random variables, rather than fixed quantities. However, the model (2.1) is generally interpreted as one of regression, and in that context the design variables x_i would be conditioned upon, even if they had a random origin. The study of (2.1), when the x_i's are treated as random variables, is generally viewed as correlation analysis rather than regression.

Depending on the performance of the pilot estimator in the true model, it could be the case that the final estimator $\bar\beta$ actually performs less well than the pilot estimator. For example, this would tend to be the case if the pilot estimator were the standard least-squares estimator and the true β did not have any components close to zero. This property is shared by the method proposed by Zou (2006), and also arises in related problems in inference.

2.3. Empirical choice of m

It will be clear from both our numerical work and our theoretical analysis (see e.g., Theorem 3 in Section 3) that performance of the m-out-of-n bootstrap in the present problem is quite insensitive to choice of m. This fact makes it inherently difficult to choose m optimally, and so we instead opt for a method which is intuitively reasonable and selects m within the range of values where we know the method enjoys good performance.

Arguably the most impressive feature of the lasso is its ability to identify components of the parameter vector β^0 that equal zero, or are close to zero; and either eliminate them from the model or give them relatively little weight in the estimator. Our approach to an empirical choice of m is founded on this aspect of the problem. We suggest an algorithm that has the following steps. (a) Identify a small number of subsets of the components of β that might be zero. (b) Introduce a second bootstrap method, based on the subsets chosen in (a). (c) Determine the values of m that give minimum mean squared error in those settings, and select the final value of m using this information.

First we describe step (a) of the algorithm. As in Section 2.2, let $\tilde\beta$ denote a pilot estimator of β^0. Rank the absolute values of the components of $\tilde\beta$ as $|\tilde\beta_{(1)}| \leq \cdots \leq |\tilde\beta_{(p)}|$. Given a candidate value, q say, for the number of integers j for which $\beta_j^0 = 0$, determine these to be the indices of the components $\tilde\beta_{(1)}, \ldots, \tilde\beta_{(q)}$. Replace these components in $\tilde\beta$ by zero, and let $\check\beta$ denote the vector that results. (Thus, q of the components of $\check\beta$ are zero, and the other $p-q$ components are identical to their counterparts for $\tilde\beta$.)

Step (b) has the following form. Generate new data pairs $(x_i, Y_i^\dagger)$, for $1 \leq i \leq n$, where $Y_i^\dagger = \bar Y + x_i^{\mathrm{T}} \check\beta + \epsilon_i^\dagger$, with the design points $x_1, \ldots, x_n$ identical to those in the original dataset $\mathcal{Z} = \{(x_1, Y_1), \ldots, (x_n, Y_n)\}$, and with the bootstrap errors $\epsilon_i^\dagger$ either computed as were the variables ϵ_i^* in Section 2.2, or generated more simply, for example as independent Normal $\mathrm{N}(0, \hat\sigma^2)$ variates where $\hat\sigma^2 = n^{-1} \sum_i \tilde\epsilon_i^2$.

Step (c) has seven sub-steps, as follows. (i) Replace the original dataset $\mathcal{Z}$ by $\mathcal{Z}^\dagger = \{(x_1, Y_1^\dagger), \ldots, (x_n, Y_n^\dagger)\}$. (ii) Apply the bootstrap algorithm suggested in Section 2.2 to $\mathcal{Z}^\dagger$, for a particular value of m, obtaining a version $\hat\lambda^\dagger$ of the weight vector λ at (2.5). (iii) Take $\bar\beta^\dagger = (\bar\beta_1^\dagger, \ldots, \bar\beta_p^\dagger)^{\mathrm{T}}$ to be the minimizer of

$$S_{\hat\lambda^\dagger}^\dagger(\beta) = \sum_{i=1}^{n} \left(Y_i^\dagger - \bar Y^\dagger - x_i^{\mathrm{T}}\beta\right)^2 + \sum_{j=1}^{p} \hat\lambda_j^\dagger \, |\beta_j| \, ,$$

where $\bar Y^\dagger = n^{-1} \sum_i Y_i^\dagger$. (iv) Compute $\|\bar\beta^\dagger - \check\beta\|^2$. (v) Average $\|\bar\beta^\dagger - \check\beta\|^2$ over many simulated versions of $\mathcal{Z}^\dagger$, thereby obtaining $\hat s_q(m) = E(\|\bar\beta^\dagger - \check\beta\|^2 \,|\, \mathcal{Z})$, which is a bootstrap estimator of $s_q(m) = E(\|\bar\beta - \beta^0\|^2)$ in the case where just q of the components of β^0 vanish. (vi) Choose $m = \widehat m_q$ to minimize $\hat s_q(m)$. (vii) Take the final value of m to be the average of values of $\widehat m_q$ for integers

q which we believe, perhaps after formal tests of statistical significance, to be approximations to the number of zero components β_j^0 in β^0.

3. Theoretical Properties

3.1. Theoretical properties under local perturbations

To adequately model the variety of ways in which the components of β can vary, especially in the context of bootstrap methods, we initially take the true value of β, $\beta^0 = \beta^{n0}$ say, to depend on n and assume that, for an integer q with $0 \le q \le p$, and finite constants $\gamma_1, \dots, \gamma_q$,

$$\begin{array}{l}\text{for } 1 \le j \le q,\ n^{1/2}\,\beta_j^0 \to \gamma_j, \text{ and for } q+1 \le j \le p,\ s_j \equiv \operatorname{sgn}(\beta_j^0) \\ \text{does not depend on } n, \text{ and } n^{1/2}\,|\beta_j^0| \to \infty.\end{array} \tag{3.1}$$

That is, we reorder of the components of β^0 so that the first q are "small," in fact perturbations of zero on the scale of $n^{-1/2}$, and the remaining $p-q$ are "large." We also impose minor conditions on the design points x_i:

$$n^{-\frac{1}{2}} \max_{1 \le i \le n} \|x_i\| \to 0 \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^{n} x_i\, x_i^{\mathrm{T}} \to \Sigma\,, \tag{3.2}$$

where Σ is a $p \times p$ matrix and

$$\Sigma \quad \text{is positive definite}\,. \tag{3.3}$$

Given a vector $\lambda^0 = (\lambda_1^0, \dots, \lambda_p^0)^{\mathrm{T}}$ with nonnegative, finite components, let Z denote a random p-vector with the normal $\mathrm{N}(0, \sigma^2\,\Sigma)$ distribution, and define

$$V(u \,|\, \lambda^0) = u^{\mathrm{T}}\,\Sigma\,u - 2\,u^{\mathrm{T}} Z + \sum_{j=1}^{q} \lambda_j^0\,(|\gamma_j + u_j| - |\gamma_j|) + \sum_{j=q+1}^{p} \lambda_j^0\, s_j\, u_j\,.$$

The methods leading to Theorem 2 of Knight and Fu (2000) can be used to derive the following result.

Theorem 1. *If* (3.1)–(3.3) *hold and* $n^{-1/2}\,\lambda \to \lambda^0$*, then*

$$n^{\frac{1}{2}}(\hat\beta - \beta^0) \to U(\lambda^0) \equiv \operatorname{argmin}_u V(u \,|\, \lambda^0)$$

in distribution, and $E\big\|n^{1/2}\,(\hat\beta - \beta^0)\big\|^2 \to E\|U(\lambda^0)\|^2$.

The optimal value of λ^0, λ^0_{opt} say, minimizes $E\|U(\lambda^0)\|^2$. If $\gamma_j = 0$ for $1 \le j \le q$, then the jth component of λ^0_{opt} equals infinity if $1 \le j \le q$ and equals zero otherwise; see the lemma in Section 5.1. Therefore, if β^0 were fixed then an asymptotically optimal estimator of β^0 would have $\hat\beta_j = o_p(n^{-1/2})$ in cases where $\beta_j^0 = 0$, and would have the other components $\hat{\beta}_k$ determined by, in effect, ordinary least-squares restricted to indices k for which $\beta_k^0 \neq 0$.

However, the situation changes substantially if, in (3.1), one or more of the γ_j's, for $1 \leq j \leq q$, is nonzero. For such a value of j the corresponding λ_j^0 is no longer infinite, and so the distribution of $U(\lambda^0)$ changes. Therefore a perturbation, or inaccuracy, of order $n^{-1/2}$ in the value of β_j^0 for sample size n, relative to the value of β_j^0 in the limit as $n \to \infty$, can significantly alter the asymptotic distribution of $\hat{\beta}$. This is the reason the standard bootstrap fails to consistently estimate the optimal λ; further details are given below.

3.2. Why the standard bootstrap fails, and the m-out-of-n bootstrap works

Suppose that the true value, β_j^0, of β_j is zero for $1 \leq j \leq q$ (and for $n \geq 1$), and equals a fixed, nonzero constant for $q + 1 \leq j \leq p$. If $1 \leq j \leq q$ then interpreting Theorem 1 in the case of the m-out-of-n bootstrap involves, in the resampling step, replacing β_j^0 by its pilot estimator, $\tilde{\beta}_j$ (see Section 2.2), which is generally in error by terms of order $n^{-1/2}$. That is, $\tilde{\beta}_j = \beta_j^0 + n^{-1/2}\zeta_j$, where ζ_j is an asymptotically normally distributed estimation error with, in the large-sample limit, zero mean and finite, nonzero variance. When resampling from this empirical approximation to the true model we would take γ_j, in (3.1), to be the large-sample limit of ζ_j. In the resampling step we condition on the data, and in particular we hold ζ_j fixed. The presence of this perturbation is the reason for the failure of the standard bootstrap; as noted in the last paragraph of the previous subsection, it alters the limiting distribution.

However, if we use instead the m-out-of-n bootstrap then, while the identity $\tilde{\beta}_j = \beta_j^0 + n^{-1/2}\zeta_j$ is still appropriate (since $\tilde{\beta}_j$ is computed from a sample of size n), Theorem 1 should now be interpreted on a scale of $m^{-1/2}$ rather than $n^{-1/2}$. Therefore, attention focuses now on $m^{1/2}\beta_j^0$, not on $n^{1/2}\beta_j^0$ (see (3.1)), with β_j^0 replaced by $\tilde{\beta}_j = \beta_j^0 + n^{-1/2}\zeta_j$. When the true value of β_j equals 0, which is the case of critical interest (see the first sentence of this subsection), $m^{1/2}\tilde{\beta}_j = m^{1/2}n^{-1/2}\zeta_j \to 0$, provided $m/n \to 0$. Therefore the effective value of γ_j, in (3.1), is now zero; the asymptotically optimal value of λ_j^0 is infinity; and the m-out-of-n bootstrap results in an asymptotically optimal choice of λ. Details are given in Section 3.3.

These arguments apply only in cases where $1 \leq j \leq q$. If no values of β_j^0 vanish then the standard least-squares estimator is asymptotically optimal, in the sense of minimizing $E\|\hat{\beta} - \beta^0\|^2$, and the standard bootstrap, and the m-out-of-n bootstrap, both produce estimators of λ which satisfy $\hat{\lambda}/n^{1/2} \to 0$ and so are asymptotically equivalent to the least-squares estimator.

3.3. Theory for the m-out-of-n bootstrap

We first set up, and solve, a non-bootstrap version of the problem in terms of a triangular array of sub-problems, indexed by "time" n. This makes it possible to apply the result to the bootstrap case. Note that in the bootstrap world, even the error distributions vary from one value of n to another.

Assume that at time n we observe data generated under the model $Y_{ni} = \beta_{n0} + x_{ni}^{\mathrm{T}}\,\beta^{n0} + \epsilon_{ni}$, where $\beta^{n0} = (\beta_1^{n0}, \ldots, \beta_p^{n0})^{\mathrm{T}}$, and the experimental errors $\epsilon_{n1}, \ldots, \epsilon_{np}$ are independent and identically distributed with zero mean and variance σ_n^2. Suppose too that as $n \to \infty$,

$$\max_{1\le i\le n} \|x_{ni}\| = o\big(n^{\frac{1}{2}}\big), \quad \sum_{i=1}^{n} x_{ni} = 0, \quad \Sigma_n \equiv \frac{1}{n}\sum_{i=1}^{n} x_{ni}\, x_{ni}^{\mathrm{T}} \to \Sigma, \quad \sigma_n^2 \to \sigma^2, \tag{3.4}$$

$$\Sigma \text{ is a positive-definite } p \times p \text{ matrix, } 0 < \sigma^2 < \infty\,, \tag{3.5}$$

$$\lim_{c\to\infty}\, \limsup_{n\to\infty}\, E\big\{\epsilon_{n1}^2\, I(|\epsilon_{n1}| > c)\big\} = 0\,, \tag{3.6}$$

$$a_n \equiv n^{\frac{1}{2}} \max_{1\le j\le q} |\beta_j^{n0}| \to 0\,, \quad b_n \equiv n^{\frac{1}{2}} \min_{q+1\le j\le p} |\beta_j^{n0}| \to \infty\,, \tag{3.7}$$

where $0 \le q \le p$. Condition (3.4) asserts that the covariance matrix Σ_n, and variance σ_n^2, are asymptotically nondegenerate and constant; (3.6) is a standard Lindeberg-type condition for the errors; and (3.7) implies that the indices j are ordered in such a way that the first q are small on a scale of $n^{-1/2}$, and the remainder are large on that scale. In this new notation, and defining $\bar{Y}_n = n^{-1}\sum_i Y_{ni}$, we have

$$S_\lambda(\beta) = \sum_{i=1}^{n} \big(Y_{ni} - \bar{Y}_n - x_{ni}^{\mathrm{T}}\beta\big)^2 + \sum_{j=1}^{p} \lambda_j\, |\beta_j|\,.$$

Take $\lambda = \lambda_{n\,\mathrm{opt}}$, the value of λ that minimizes $E\|\hat\beta - \beta^{n0}\|^2$ when $\hat\beta$ is selected to minimize $S_\lambda(\beta)$. Denote by $\hat\beta_{\mathrm{opt}}$ this version of $\hat\beta$. If $\lambda_{n\,\mathrm{opt}}$ is not uniquely defined, choose a value that minimizes $\sum_j (\lambda_{n\,\mathrm{opt}})_j^2$, where $(\lambda_{n\,\mathrm{opt}})_j^2$ denotes the square of the jth component of $\lambda_{n\,\mathrm{opt}}$. Let Σ^0 denote the $(p-q)\times(p-q)$ matrix obtained by deleting the first q rows and q columns of Σ. Write $\mathcal{S}$ for the set $\{1, \ldots, q\}$, i.e., the set of indices j such that $\lim_{n\to\infty} n^{1/2}\,|\beta_j^{n0}| \to 0$, and let $\widehat{\mathcal{S}}$ be the set of j for which $(\hat\beta_{\mathrm{opt}})_j = 0$.

Theorem 2. *Assume that* (3.4)$-$(3.7) *hold. Then,* (a) $(\lambda_{n\,\mathrm{opt}})_j/n^{1/2} \to \infty$ *for* $1 \le j \le q$, *and* $(\lambda_{n\,\mathrm{opt}})_j/n^{1/2} \to 0$ *for* $q+1 \le j \le p$; (b) $P(\widehat{\mathcal{S}} = \mathcal{S}) \to 1$ *as* $n \to \infty$; *and* (c) *the variables* $n^{1/2}\,(\hat\beta_{\mathrm{opt}} - \beta^{n0})_j$, *for* $q+1 \le j \le p$, *are asymptotically jointly normally distributed with zero means and covariances given by the respective components of* $\sigma^2\,(\Sigma^0)^{-1}$.

In summary, Theorem 2 asserts that $\hat\beta_{\text{opt}}$ is first-order equivalent to the estimator constrained by taking $\hat\beta_j = 0$ for $1 \le j \le q$, and $\hat\beta_j$ (for $q+1 \le j \le p$) equal to the jth component of the least-squares estimator obtained after eliminating $\beta_1, \ldots, \beta_q$ from the model. Together, (b) and (c) demonstrate asymptotic optimality of the estimator $\hat\beta_{\text{opt}}$; property (b) is sometimes referred to as consistency.

Next we state a version of Theorem 2 in the bootstrap case, under (2.1) rather than the triangular array of (3.4)−(3.7). The centering assumption (2.2) implies that the design variables x_i are notionally recentered for each n, and it is advantageous here to be specific about the manner of centering. We suppose that

$$\begin{array}{l} x_i\,(= x_{ni}) = z_i - \bar z_n, \text{ where } \bar z_n = n^{-1}\sum_{i\le n} z_i \text{ and } z_1, z_2, \ldots \text{ is} \\ \text{a sequence of } p\text{-vectors for which } \|z_i\| \text{ is uniformly bounded and} \\ \Sigma_n \equiv n^{-1}\sum_{i\le n}(z_i - \bar z_n)(z_i - \bar z_n)^{\mathrm T} \to \Sigma, \text{ with } \Sigma \text{ denoting a positive-} \\ \text{definite } p\times p \text{ matrix.} \end{array} \tag{3.8}$$

In the bootstrap algorithm, take the pilot estimator $\tilde\beta$ to be the ordinary least-squares estimator. Recall the definition of $\hat\lambda = (\hat\lambda_1, \ldots, \hat\lambda_p)^{\mathrm T}$ at (2.5), and that $\bar\beta$ is chosen to minimize $S_{\hat\lambda}(\beta)$, with $S_\lambda(\beta)$ given by (2.3). Let $\mathcal{T} = \{1, \ldots, q\}$, and $\widehat{\mathcal{T}}$ be the set of j for which $\bar\beta_j = 0$.

Theorem 3. *Assume that, in the model at* (2.1), $\beta_j = 0$ *for* $1 \le j \le q$, $\beta_j \ne 0$ *for* $q+1 \le j \le p$, *the errors* ϵ_i *are independent and identically distributed with zero mean and finite variance* σ^2, *the design variables* x_i *satisfy* (3.8), *and the resample size* $m = m(n)$ *satisfies* $m = O\{n/(\log n)^{1+\eta}\}$ *and* $m \to \infty$ *for some* $\eta > 0$. *Then,* (*a*) *with probability 1,* $\hat\lambda_j/n^{1/2} \to \infty$ *for* $1 \le j \le q$ *and* $\hat\lambda_j/n^{1/2} \to 0$ *for* $q+1 \le j \le p$; (*b*) $P(\widehat{\mathcal{T}} = \mathcal{T}) \to 1$ *as* $n \to \infty$; *and* (*c*) *the variables* $n^{1/2}(\bar\beta - \beta^0)_j$, *for* $q+1 \le j \le p$, *are asymptotically jointly normally distributed with zero means and covariances given by the respective components of* $\sigma^2(\Sigma^0)^{-1}$.

Parts (b) and (c) of the theorem demonstrate that $\bar\beta$ has the oracle property; see (2.4).

3.4. The case of very large p

Versions of the theory given above can be developed in the case where $p, q \to \infty$ and $p - q$ is fixed. However, the maximum rate of divergence permitted for p seems to depend on the "generalized parameters" of the model, for example on the tail weight assumed of the distribution of ϵ, and on the distance of Σ_n from a nonsingular matrix. In addition, the biased bootstrap method becomes more labour-intensive as p increases. The competing method of Zou (2006) also suffers difficulties for large p.

4. Numerical Properties

4.1. Implementation of m-out-of-n bootstrap algorithms

We first note that one cannot use the LARS algorithm to minimize (2.3) since the penalties λ_j are different for different coefficients. Minimization of (2.3) is equivalent to the following problem:

$$\begin{aligned}&\text{minimize} \sum_{i=1}^{n} \bigl\{Y_i - \bar{Y} - x_i^{\mathrm{T}}(\beta^+ - \beta^-)\bigr\}^2 + \sum_{j=1}^{p} \lambda_j\,(\beta_j^+ + \beta_j^-) \quad \text{over } \beta^+ \\ &\quad = (\beta_1^+, \dots, \beta_p^+) \text{ and } \beta^- = (\beta_1^-, \dots, \beta_p^-), \text{ subject to } \beta_j^+, \beta_j^- \geq 0 \text{ for all } j.\end{aligned} \tag{4.1}$$

The problem (4.1) involves quadratic programming with inequality constraints, and thus can be solved by a standard technique such as the gradient projection method; see, for example, Nocedal and Wright (2006, Sec. 16.7).

Minimization of $\widehat{\mathrm{MSE}}(\lambda)$, at (2.5), on a p-dimensional grid of λ is computationally expensive when p is large. To reduce the computational burden, we suggest an iterative algorithm that has the following steps. (a) Initialize using $\lambda^{(0)} = (\lambda_1^{(0)}, \dots, \lambda_p^{(0)})^{\mathrm{T}}$. (b) For $1 \leq j \leq p$, update $\lambda_j^{(0)}$ by

$$\lambda_j^{(1)} = \mathrm{argmin}_{\lambda_j}\, \widehat{\mathrm{MSE}}\bigl(\lambda_1^{(1)}, \dots, \lambda_{j-1}^{(1)}, \lambda_j, \lambda_{j+1}^{(0)}, \dots, \lambda_p^{(0)}\bigr). \tag{4.2}$$

(c) For $k \geq 1$, repeat (b) to get $\lambda^{(k+1)}$ from $\lambda^{(k)}$ until $|\widehat{\mathrm{MSE}}(\lambda^{(k+1)}) - \widehat{\mathrm{MSE}}(\lambda^{(k)})|$ is sufficiently small. (d) Take $\tilde{\lambda}$ to be the limit of the iteration.

For the initial values in the above iteration, one may use $\lambda_j^{(0)} = c_m/|\tilde{\beta}_j|$ for some $c_m > 0$, where $\tilde{\beta}$ is the standard least-squares estimator. This initial choice corresponds to Zou's (2006) adaptive lasso penalty with $\gamma = 1$, and is closely related to the nonnegative garotte of Breiman (1995). In the simulation study we used $c_m = n^{-1/4}\, m^{1/2}$. This was based on the fact that the penalties $\lambda_j = d_n/|\tilde{\beta}_j|$ yield the oracle property for the lasso estimator if $d_n/n^{1/2} \to 0$ and $d_n \to \infty$ as $n \to \infty$; see Theorem 2 of Zou (2006). The choice $c_m = n^{-1/4}\, m^{1/2}$ corresponds to $d_n = n^{1/4}$, due to the normalization $(n/m)^{1/2}$ in (2.5).

4.2. Simulation study

We compared the finite-sample performance of the bootstrap penalty choice with Zou's (2006) approach for the lasso. In our comparison we also included the SCAD method of Fan and Li (2001).

We computed our proposed estimator $\bar{\beta}$ using the method described in Section 4.1. We used the LARS algorithm to compute Zou's (2006) adaptive lasso, which is given by $\hat{\beta}_{\mathrm{Zou},j} = b_j \, |\tilde{\beta}_j|^\gamma$, where $\beta = (b_1, \ldots, b_p)$ minimizes

$$\sum_{i=1}^{n} \left\{ Y_i - \bar{Y} - \sum_{j=1}^{p} (|\tilde{\beta}_j|^\gamma x_{ij}) \, \beta_j \right\}^2 + \zeta \sum_{j=1}^{p} \, |\beta_j| \, .$$

The tuning parameter (γ, ζ) was chosen by five-fold cross-validation. To implement the SCAD method we applied the LQA algorithm of Fan and Li (2001). This involves Newton-Raphson iteration based on a local quadratic approximation of the SCAD penalty. Specifically, let $\hat{\beta}^{(k)}_{\mathrm{SCAD}}$ be the estimate from the kth iteration. In the $(k+1)$st iteration, the algorithm sets $\hat{\beta}^{(k+1)}_{\mathrm{SCAD},j} = 0$ if $\hat{\beta}^{(k)}_{\mathrm{SCAD},j}$ is very close to zero. Let $\mathcal{A} \equiv \mathcal{A}^{(k)}$ be the set of indices for the other components. Then, the algorithm takes

$$\hat{\beta}^{(k+1)}_{\mathrm{SCAD},\mathcal{A}} = \left\{ X_{\mathcal{A}}^{\mathrm{T}} X_{\mathcal{A}} + n \, D_\omega(\hat{\beta}^{(k)}_{\mathrm{SCAD}}) \right\}^{-1} X_{\mathcal{A}}^{\mathrm{T}} Y \, ,$$

where $X_{\mathcal{A}}$ denotes the design matrix consisting of the columns that correspond to those $j \in \mathcal{A}$, $D_\omega(\theta) = \mathrm{diag}\{p'_\omega(|\theta_j|)/|\theta_j|\}_{j \in \mathcal{A}}$ and

$$p'_\omega(\theta) = \omega \left\{ I(\theta \le \omega) + \frac{(a\omega - \theta)_+}{(a-1)\,\omega} \, I(\theta > \omega) \right\} , \quad \text{where} \quad a > 2 \, .$$

In the simulation we set

$$\hat{\beta}^{(k+1)}_{\mathrm{SCAD},j} = 0 \quad \text{if} \quad \big| \hat{\beta}^{(k)}_{\mathrm{SCAD},j} \big| < \delta \sum_{i=1}^{p} \big| \hat{\beta}^{(k)}_{\mathrm{SCAD},i} \big| \tag{4.3}$$

for a small positive number δ. We took the standard least-squares estimator $\tilde{\beta}$ as the initial value $\hat{\beta}^{(0)}_{\mathrm{SCAD}}$, and used five-fold cross-validation to choose the tuning parameter (a, ω).

We generated the values of the predictor $x_i = (x_{i1}, \ldots, x_{i8})^{\mathrm{T}}$ from the normal $\mathrm{N}_8(0, \Sigma)$ distribution, where $\Sigma = (0.5^{|i-j|})$. We set $\sigma = 1$ and 3. The sample sizes were $n = 50$, 100. For the values of the true parameters we considered the following two models:

Model 1: $\beta^0 = (3, 1.5, 0, 0, 2, 0, 0, 0)$.

Model 2: $\beta^0_j = 1.5$ for $1 \le j \le 3$ and $\beta^0_j = 0.3$ for $4 \le j \le 8$.

Note that the first model was also considered by Fan and Li (2001) and Zou (2006). Comparison was made in terms of the numbers of correct and incorrect identifications of zero parameters, as well as the prediction error PE $= E(Y -$

$x^{\mathrm{T}}\bar{\beta})^2$, where the expectation was taken with respect to the test observation (x, Y) only. Since

$$\mathrm{PE} = \sigma^2 + (\bar{\beta} - \beta^0)^{\mathrm{T}} E(xx^{\mathrm{T}})(\bar{\beta} - \beta^0),$$

we used $\mathrm{ME} = (\bar{\beta} - \beta^0)^{\mathrm{T}} E(xx^{\mathrm{T}})\,(\bar{\beta} - \beta^0)$ as a measure of performance; here, ME stands for "model error."

One hundred bootstrap samples were used to compute $\widehat{\mathrm{MSE}}(\lambda)$. For a given value of λ, it took about 2 seconds to obtain $\widehat{\mathrm{MSE}}(\lambda)$ when $n = 50$ and $m = 40$ in our current computing environment with CPU: intel(R) Core2 Duo 1.86GHz and RAM: 1GB. Updating a single λ_j as at (4.2) took roughly 7.5 seconds, so that it took about 1 minute to run one whole iteration for updating $\lambda_j, j = 1, \ldots, 8$. The tolerance value in the iteration for selecting $\tilde{\lambda}$ as described in Section 4.1 was set at 10^{-6}. We found that in most cases the algorithm converged in five to ten iterations. Thus, the average computing time to compute $\hat{\lambda}$ at (2.5) was about 6 minutes.

Tables 1 and 2 summarize the results of the simulation for several choices of the subsample size m for the bootstrap lasso, and of the cut-off value δ, at (4.3), for the SCAD method. The tables show the median value of relative model error (MRME) with respect to the least-squares estimator, as well as the average numbers of correct and incorrect identifications of zero coefficients. The measure MRME was also used by Fan and Li (2001). The bootstrap lasso worked better for smaller m in Model 1, and for larger m in Model 2. This was to be expected since a small m, if not too small, would lead to choosing a penalty λ close to its optimum when there were zero coefficients. Note too that subsampling was not so crucial to performance of the bootstrap method when there was no zero coefficient; see the discussion in Section 3.2. In contrast, the SCAD method performed well for relatively large δ when there were zero coefficients. This was also expected since, for larger values of δ, it was more likely that the estimates of the zero coefficients were equal to zero.

Comparing, in the context of Model 1, Zou's adaptive lasso, the bootstrap lasso and the SCAD approach, using tuning-parameter values that gave optimal performance of the respective methods, we found that the bootstrap lasso performed best when the noise level was high, while the SCAD method was best when the noise level was low. Zou's adaptive lasso did not do as well, but its performance did not vary much for different noise levels. On the other hand in the case of Model 2, Zou's performed best. In this setting the bootstrap lasso was again better (worse) than the SCAD technique when the noise level was high (low), again when tuning-parameter values were chosen so as to give optimum performance.

Table 1. Comparison of the Methods Based on 100 Replications for Model 1.

Method		MRME(%)	*Avg. No. of* 0 *Coefficients* Correct	 Incorrect
$(n = 50, \sigma = 1)$				
Bootstrap LASSO	$m = 10$	49.80	4.66	0.01
	$m = 20$	48.57	4.46	0
	$m = 40$	58.15	4.16	0
SCAD	$\delta = 0.01$	64.30	3.95	0
	$\delta = 0.03$	55.78	4.49	0
	$\delta = 0.06$	34.47	4.97	0
Zou's LASSO		73.63	2.68	0
$(n = 50, \sigma = 3)$				
Bootstrap LASSO	$m = 10$	65.30	4.03	0.19
	$m = 20$	64.92	3.63	0.07
	$m = 40$	75.30	3.01	0.05
SCAD	$\delta = 0.01$	86.32	3.10	0.07
	$\delta = 0.03$	81.71	3.57	0.08
	$\delta = 0.06$	77.52	3.95	0.08
Zou's LASSO		71.15	2.89	0.02
$(n = 100, \sigma = 1)$				
Bootstrap LASSO	$m = 20$	44.04	4.79	0
	$m = 40$	46.62	4.59	0
	$m = 80$	55.48	4.22	0
SCAD	$\delta = 0.01$	56.12	4.40	0
	$\delta = 0.03$	46.64	4.80	0
	$\delta = 0.06$	37.93	5	0
Zou's LASSO		72.89	2.85	0
$(n = 100, \sigma = 3)$				
Bootstrap LASSO	$m = 20$	51.36	4.28	0.01
	$m = 40$	62.93	3.75	0.01
	$m = 80$	72.84	3.26	0
SCAD	$\delta = 0.01$	65.84	3.84	0
	$\delta = 0.03$	61.83	4.14	0
	$\delta = 0.06$	65.23	4.41	0
Zou's LASSO		73.28	2.95	0

Table 2. Comparison of the Methods Based on 100 Replications for Model 2.

Method		MRME(%)	*Avg. No. of* 0 *Coefficients* Correct	 Incorrect
$(n = 50, \sigma = 1)$				
Bootstrap LASSO	$m = 10$	191.37	-	2.12
	$m = 20$	143.58	-	1.60
	$m = 40$	126.06	-	1.29
SCAD	$\delta = 0.01$	103.92	-	1.05
	$\delta = 0.03$	126.05	-	1.53
	$\delta = 0.06$	269.59	-	3.17
Zou's LASSO		100.00	-	0.30
$(n = 50, \sigma = 3)$				
Bootstrap LASSO	$m = 10$	108.34	-	3.52
	$m = 20$	100.17	-	3.17
	$m = 40$	89.56	-	2.67
SCAD	$\delta = 0.01$	100.00	-	2.37
	$\delta = 0.03$	97.94	-	2.76
	$\delta = 0.06$	100.56	-	3.39
Zou's LASSO		83.98	-	2.36
$(n = 100, \sigma = 1)$				
Bootstrap LASSO	$m = 20$	185.43	-	1.11
	$m = 40$	160.35	-	0.69
	$m = 80$	131.81	-	0.54
SCAD	$\delta = 0.01$	100.00	-	0.49
	$\delta = 0.03$	128.73	-	0.96
	$\delta = 0.06$	516.58	-	3.47
Zou's LASSO		100.00	-	0.06
$(n = 100, \sigma = 3)$				
Bootstrap LASSO	$m = 20$	123.10	-	3.18
	$m = 40$	108.05	-	2.84
	$m = 80$	98.54	-	2.38
SCAD	$\delta = 0.01$	100.19	-	1.96
	$\delta = 0.03$	103.35	-	2.49
	$\delta = 0.06$	115.77	-	3.16
Zou's LASSO		87.68	-	1.36

Table 3. Performance of Bootstrap LASSO with Empirical Choice of m.

	Noise		*Avg. No. of* 0 *Coefficients*	
Model	Level	MRME(%)	Correct	Incorrect
Model 1	$\sigma = 1$	50.36	4.33	0
	$\sigma = 3$	70.90	3.48	0.08
Model 2	$\sigma = 1$	125.65	-	1.25
	$\sigma = 3$	89.86	-	2.71

Note: Based on 100 replications with sample size $n = 50$.

From the tables we see that performance of the SCAD method depended very much on choice of the cut-off value, δ, especially when the noise level was low. For example, the value of MRME ranged from 100 to 516 when $n = 100$ and $\sigma = 1$, in the case of Model 2. One unpleasant feature of the LQA algorithm, for the SCAD method, was that it forced coefficients to be artificially zero. Furthermore, as noted by Fan and Li (2001), once a coefficient was shrunken to zero in the iteration, it stayed at that value. No guidelines have been suggested for choosing the cut-off value. Performance of the bootstrap lasso also had a degree of dependence on subsample size, as is usually the case for m-out-of-n bootstrap procedures. In contrast to the SCAD method, however, we have a rule for choosing m empirically, as suggested in Section 2.3.

Indeed, we investigated performance of the bootstrap lasso when the subsample size, m, was chosen empirically according to the algorithm described in Section 2.3. To determine q, the number of zero components of $\check{\beta}$ in the bootstrap population for generating bootstrap samples $\{(x_i, Y_i^{\dagger})\}$, we first ranked the absolute values of the components of $\tilde{\beta}$ as $|\tilde{\beta}_{(1)}| \le \cdots \le |\tilde{\beta}_{(p)}|$. For $j = 1, \ldots, p$, we conducted a series of F-tests for the hypotheses $\mathrm{H}_j : \beta^0_{(1)} = \cdots = \beta^0_{(j)} = 0$. For two given levels of significance α_1 and α_2 ($\alpha_1 > \alpha_2$), we took q_1 and q_2, respectively, to be the largest indices j such that H_j was accepted, and we took the values of q to be those in the interval $[q_1, q_2]$. In our simulation, we chose $\alpha_1 = 0.05$ and $\alpha_2 = 0.01$.

Table 3 gives simulation results for the bootstrap lasso when the subsample size, m, was chosen empirically according to the algorithm described in Section 2.3 and in the previous paragraph. From the table we see that the empirical choice of m gave nearly the same performance as the one that the bootstrap lasso achieved with optimally chosen tuning-parameter values. The resulting bootstrap lasso was better than the SCAD method with optimally chosen δ, when the noise level was high.

5. Technical Arguments

5.1. Proof of Theorem 2

Step 1 : *Minimum mean squared error in the asymptotic limit.* Let $\nu_1, \ldots, \nu_q$ denote nonnegative constants, let $\mu_{q+1}, \ldots, \mu_p$ be arbitrary real numbers, let μ

represent the p vector of which the first q components equal 0 and the next $p-q$ equal $\mu_{q+1},\ldots,\mu_p$, and write ξ for a general p-vector. Let Σ denote a $p\times p$, positive-definite matrix, let Σ^0 be as in Theorem 2, let $\sigma>0$, and write W for a random p-vector having the normal $\mathrm{N}(0,\sigma^2\,\Sigma)$ distribution. Define $\xi=\Xi(\mu,\nu)$, written below simply as Ξ, to be a random vector that minimizes

$$\xi^{\mathrm{T}}\Sigma\,\xi-2\,\xi^{\mathrm{T}}W+\sum_{j=1}^{q}\nu_j\,|\xi_j|+\mu^{\mathrm{T}}\xi\,. \tag{5.1}$$

Lemma. *The values of $\nu_1,\ldots,\nu_q$ and μ that minimize $E\|\Xi\|^2$ are $\nu_j=\infty$, for $1\le j\le q$, and $\mu=0$; the minimum is unique. Moreover, a vector Ξ that minimizes $E\|\Xi\|^2$ has $\Xi_j=0$ almost surely, for $1\le j\le q$, and has its other $p-q$ components jointly normally distributed with zero means and covariances given by the respective components of $\sigma^2\,(\Sigma^0)^{-1}$.*

Taking each $\nu_j=\infty$ and $\mu=0$ ensures that Ξ is just the minimizer of $\xi^{\mathrm{T}}\Sigma\,\xi-2\,\xi^{\mathrm{T}}W$, after the first q components of ξ are constrained to equal to 0.

Proof of Lemma. Without loss of generality, $\sigma=1$. Consider the problem of estimating $\beta^0=(\beta_1^0,\ldots,\beta_p^0)^{\mathrm{T}}$ in the regression model $Y_i=x_i^{\mathrm{T}}\,\beta^0+\epsilon_i$, where the errors ϵ_i are independent and identically distributed as Normal $\mathrm{N}(0,1)$, the design variables x_i satisfy $\Sigma_n\equiv n^{-1}\sum_i x_ix_i^{\mathrm{T}}\to\Sigma$, we know that $\beta_1^0=\cdots=\beta_q^0=0$, and we know only the signs of $\beta_{q+1}^0,\ldots,\beta_p^0$, all of which are nonzero. Since the errors are Gaussian then the least-squares estimator of β^0, $\tilde\beta_{\mathrm{ls}}$ say, constructed after equating to zero the estimators of $\beta_1^0,\ldots,\beta_q^0$, has minimum asymptotic variance and minimum asymptotic mean squared error.

Defining $\xi=(\xi_1,\ldots,\xi_p)^{\mathrm{T}}=n^{1/2}(\beta-\beta^0)$, $\tilde\xi_{\mathrm{ls}}=n^{1/2}\,(\tilde\beta_{\mathrm{ls}}-\beta^0)$ can be shown to equal the minimizer of $\xi^{\mathrm{T}}\Sigma_n\,\xi-2\,\xi^{\mathrm{T}}W_n$, with $\xi_1,\ldots,\xi_q$ constrained to equal 0, where $W_n=n^{-1/2}\sum_i x_i\,\epsilon_i$ and so is distributed as normal $\mathrm{N}(0,\Sigma_n)$. The asymptotic distribution of $\tilde\xi_{\mathrm{ls}}$ is that of the minimizer, $\tilde\xi_{\mathrm{as}}$ say, of $\xi^{\mathrm{T}}\Sigma\,\xi-2\,\xi^{\mathrm{T}}W$, constrained to have $\xi_1=\cdots=\xi_q=0$. This is the same as the limiting distribution of the quantity that minimizes the function of ξ at (5.1), provided we take $\nu_j=\infty$ for $1\le j\le q$ (this imposes the constraint $\xi_1=\cdots=\xi_q=0$) and $\mu=0$.

A second, competing estimator of β^0, the asymptotic distribution of which is the distribution of the minimizer of (5.1), is the vector $\beta=\tilde\beta$ that minimizes

$$\sum_{i=1}^{n}\left(Y_i-x_i^{\mathrm{T}}\beta\right)^2+n^{\frac12}\sum_{j=1}^{q}\nu_j\,|\beta_j|+n^{\frac12}\sum_{j=q+1}^{p}\alpha_j\,|\beta_j|\,,$$

where $\alpha_j=\mu_j\,\mathrm{sgn}(\beta_j^0)$. Minimum variance of $\tilde\xi_{\mathrm{ls}}$, and hence also the fact that $E\|\Xi\|^2$ is minimized by taking each $\nu_j=\infty$ and $\mu=0$, follows from the minimum-variance property noted at the end of the first paragraph of this proof. Uniqueness of the minimum of $E\|\Xi\|^2$ follows from uniqueness of the minimum-variance property. The correctness of the second sentence in the lemma follows from the definition of Ξ.

Step 2 : *Simplification of formula for* S_λ. Define $\gamma_{nj} = n^{1/2}\,\beta_j^{n0}$, $\Delta_j = n^{1/2}\,(\beta_j - n^{-1/2}\,\gamma_{nj})$, $\Delta = (\Delta_1, \ldots, \Delta_p)^{\mathrm{T}}$, $\lambda_{nj} = n^{-1/2}\,(\lambda_{n\,\mathrm{opt}})_j$, $\Sigma_n = n^{-1}\sum_i x_{ni}x_{ni}^{\mathrm{T}}$ and $Z_n = n^{-1/2}\sum_i \epsilon_{ni}\,x_{ni}$. Write R for a random variable that does not depend on β. In this notation,

$$S_{\lambda_{n\,\mathrm{opt}}}(\beta) = \Delta^{\mathrm{T}}\,\Sigma_n\Delta - 2\,Z_n^{\mathrm{T}}\Delta \\ +\sum_{j=1}^{q}\lambda_{nj}\,|\gamma_{nj}+\Delta_j| + \sum_{j=q+1}^{p}\lambda_{nj}\,(|\gamma_{nj}+\Delta_j| - |\gamma_{nj}|) + R\,. \tag{5.2}$$

Note too that, by (3.4) and (3.5),

$$\limsup_{n\to\infty} E\|Z_n\|^2 < \infty\,. \tag{5.3}$$

Observe too that the first, third and fourth parts of (3.4), and the Lindeberg-type condition (3.6), imply that

$$Z_n \;\to\; N(0, \sigma^2\,\Sigma) \;\; \text{in distribution}\,. \tag{5.4}$$

The standard least-squares estimator $\tilde\beta$, which minimizes $S_\lambda(\beta)$ when $\lambda \equiv 0$, has $n^{1/2}\,(\tilde\beta - \beta) = \Sigma_n^{-1}Z_n$, and so satisfies, for all sufficiently large n, $n\,E\|\tilde\beta - \beta\|^2 = \mathrm{tr}(\Sigma_n^{-1})\,\sigma^2 \le C_3 \equiv 2p\sigma^2/C_1$, where C_1 denotes the smallest eigenvalue of Σ. Since $\lambda = \lambda_{n\,\mathrm{opt}}$ minimizes $E\|\Delta\|^2$ then, if Δ_{opt} denotes the version of Δ computed when β is chosen to minimize $S_{\lambda_{n\,\mathrm{opt}}}(\beta)$,

$$E\|\Delta_{\mathrm{opt}}\|^2 \le C_3 \quad \text{for all sufficiently large } n\,. \tag{5.5}$$

Recalling the definition of b_n at (3.7), denote by B_n the set of β for which $n^{1/2}\,\max_{q+1\le j\le p}\,|\beta_j - \beta_j^{n0}| = \max_{q+1\le j\le p}\,|\Delta_j| \le b_n/2$. In view of (5.5),

$$P(\hat\beta_{\mathrm{opt}} \in B_n) \to 1\,. \tag{5.6}$$

Let $\lambda'_{nj} = \lambda_{nj}\,\mathrm{sgn}(\gamma_{nj})$. If $\beta \in B_n$, then by (5.2),

$$S_{\lambda_{n\,\mathrm{opt}}}(\beta) = \Delta^{\mathrm{T}}\,\Sigma_n\Delta - 2\,Z_n^{\mathrm{T}}\Delta + \sum_{j=1}^{q}\lambda_{nj}\,|\gamma_{nj}+\Delta_j| + \sum_{j=q+1}^{p}\lambda'_{nj}\,\Delta_j + R\,. \tag{5.7}$$

Step 3 : *Completion.* Assume that all the eigenvalues of Σ lie in $[C_1, C_2]$, where $0 < C_1 \le C_2 < \infty$. Choose n_1 so large that, for all $n \ge n_1$, all the eigenvalues of Σ_n lie in $[C_1/2, 2\,C_2]$. Differentiating the right-hand side of (5.7) with respect to

Δ_j, for $q+1 \leq j \leq p$, and equating to zero, so as to obtain a local minimum in $\Delta_{q+1}, \ldots, \Delta_p$ for fixed $\Delta_1, \ldots, \Delta_q$, we find that $\sum_k (\Sigma_n)_{jk}\, \Delta_k = (Z_n)_j - \lambda'_{nj}/2$. Hence, taking $\beta = \hat{\beta}_{\rm opt}$ and noting (5.6), we deduce that for all $n \geq n_1$ and each $j \in \{q+1, \ldots, p\}$,

$$2\, C_2\, \|\Delta_{\rm opt}\| + |(Z_n)_j| \geq \tfrac{1}{2}\, \lambda_{nj}\,. \tag{5.8}$$

It can be deduced from (5.3), (5.5) and (5.8) that,

$$\max_{q+1 \leq j \leq p}\ \limsup_{n\to\infty}\ \lambda_{nj} < \infty\,. \tag{5.9}$$

Next we treat λ_{nj} for $1 \leq j \leq q$. For this purpose we assume, without loss of generality, that the sequences λ_{nj} converge, as $n \to \infty$, to respective nonnegative limits $\lambda_j^{\rm lim}$, which might be either finite or infinite. (Subsequence arguments can be used if the limits are not well defined.) In view of the first part of (3.7), $\gamma_{nj} \to 0$ for $1 \leq j \leq q$, and so,

$$\text{if } 1 \leq j \leq q \text{ and } \lambda_j^{\rm lim} < \infty \text{ then } \lambda_{nj}|\gamma_{nj} + (\Delta_{\rm opt})_j| = \lambda_j^{\rm lim}|(\Delta_{\rm opt})_j| + o_p(1). \tag{5.10}$$

In the next paragraph we deal with the case $\lambda_j^{\rm lim} = \infty$.

Given $\eta > 0$, let $\mathcal{J}_n(\eta)$ denote the set of values $j \in \{1, \ldots, q\}$ such that $|(\Delta_{\rm opt})_j| > 3\eta$. Write $\Delta_{\rm opt'}$ for the version of $\Delta_{\rm opt}$ that is obtained on replacing $(\hat{\beta}_{\rm opt})_j$ by β_j^{n0} if $j \in \mathcal{J}_n(\eta)$, and leaving unchanged the other values of $(\hat{\beta}_{\rm opt})_j$. Results (5.3), (5.5) and (5.9) imply that, for $\tilde{\Delta} = \Delta_{\rm opt}$ or $\Delta_{\rm opt'}$,

$$\left|\tilde{\Delta}^{\rm T}\, \Sigma_n \tilde{\Delta} - 2\, Z_n^{\rm T} \tilde{\Delta} + \sum_{j=q+1}^{p} \lambda'_{nj}\, (\tilde{\Delta})_j\right| = O_p(1)$$

as $n \to \infty$. Hence, by (5.6) and (5.7),

$$\sum_{j=1}^{q} \lambda_{nj}\, |\gamma_{nj} + (\Delta_{\rm opt})_j| = S_{\lambda_{n\,{\rm opt}}}(\hat{\beta}_{\rm opt}) - R + O_p(1)\,, \tag{5.11}$$

$$\begin{aligned}\sum_{j\in\mathcal{J}_n(\eta)} \lambda_{nj}\, |\gamma_{nj}| + \sum_{j\notin\mathcal{J}_n(\eta)} \lambda_{nj}|\gamma_{nj} + (\Delta_{\rm opt'})_j| \\ = S_{\lambda_{n\,{\rm opt}}}(\hat{\beta}_{\rm opt'}) - R + O_p(1)\,.\end{aligned} \tag{5.12}$$

If $j \in \mathcal{J}_n(\eta)$ and n is so large that $\sup_{1\leq j\leq q} |\gamma_{nj}| \leq \eta$, then $|\gamma_{nj}| \leq (1/2)|\gamma_{nj} + (\Delta_{\rm opt})_j|$. Hence, comparing (5.11) and (5.12) we deduce that unless

$$\sum_{j\in\mathcal{J}_n(\eta)} \lambda_{nj}\, |\gamma_{nj} + (\Delta_{\rm opt})_j| = O_p(1) \tag{5.13}$$

we have a contradiction of the fact that $\hat{\beta}_{\text{opt}}$ minimizes $S_{\lambda_{n\,\text{opt}}}(\beta)$. Therefore (5.13) holds for each $\eta > 0$, which implies that,

$$\text{if } 1 \le j \le q \text{ and } \lambda_j^{\lim} = \infty \text{ then } (\Delta_{\text{opt}})_j \to 0 \text{ in probability.} \tag{5.14}$$

To complete the proof we complement the assumption that $\lambda_{nj} \to \lambda_j^{\lim}$ for $1 \le j \le q$, by supposing that this result also holds for $q+1 \le j \le p$, and that for $q+1 \le j \le p$, $\operatorname{sgn}(\gamma_{nj}) \to s_j^{\lim}$, where $s_j^{\lim}$ denotes either $+1$ or -1. The contrary case, where the limits are not well defined, can be treated using a subsequence argument. In view of (5.9),

$$\text{for } q+1 \le j \le p,\ \lambda'_{nj} \to s_j^{\lim}\,\lambda_j^{\lim} \text{ where } \lambda_j^{\lim} \text{ is finite.} \tag{5.15}$$

Without loss of generality, the first q components of β are ordered such that, for an integer $r \in \{0, \ldots, q\}$, $\lambda_j^{\lim} < \infty$ when $1 \le j \le r$, and $\lambda_j^{\lim} = \infty$ when $r+1 \le j \le q$. We know from (5.14) that if $r+1 \le j \le q$, then the asymptotic distribution of $(\Delta_{\text{opt}})_j$ is degenerate at zero. Combining this result with (5.10) and (5.15) we deduce that the components $(\Delta_{\text{opt}})_j$, for $1 \le j \le p$, are obtained, up to terms that equal $o_p(1)$, by minimizing

$$T(\beta) = \Delta^{\mathrm{T}}\,\Sigma_n\Delta - 2\,Z_n^{\mathrm{T}}\,\Delta + \sum_{j=1}^{r} \lambda_j^{\lim}\,|\Delta_j| + \sum_{j=q+1}^{p} s_j^{\lim}\,\lambda_j^{\lim}\,\Delta_j\,, \tag{5.16}$$

where it is understood that $\Delta_j = 0$ for $r+1 \le j \le q$.

Standard asymptotic arguments, noting the asymptotic normality of Z_n at (5.4), show that the limiting distribution of the minimizer of $T(\beta)$ at (5.16) is the distribution of Ξ which minimizes the quantity at (5.1), with the obvious definitions of $\nu_1, \ldots, \nu_q, \mu_{q+1}, \ldots, \mu_p$ there. Since $\lambda = \lambda_{n\,\text{opt}}$ is chosen to minimize $E\|\hat{\beta} - \beta^{n0}\|^2$, then it follows directly from the Lemma in Step 1 that $r = 0$ and $\lambda_j^{\lim} = 0$ for $q+1 \le j \le p$. This proves part (a) of Theorem 2. Part (c), together with the fact that $(\hat{\beta}_{\text{opt}})_j = o_p(n^{-1/2})$ for $1 \le j \le q$, follows from the last sentence in the lemma.

Finally we derive part (b). Part (c), and the second part of (3.7), imply that $P(\widehat{\mathcal{S}} \subseteq \mathcal{S}) \to 1$. Therefore it suffices to show that $P(\mathcal{S} \setminus \widehat{\mathcal{S}}) \to 0$. Write x_{nij} for the jth component of x_{ni}. If $(\hat{\beta}_{\text{opt}})_j \neq 0$, i.e., if $j \notin \widehat{\mathcal{S}}$, then by the Kuhn-Tucker Theorem, since $\hat{\beta}_{\text{opt}}$ gives an extremum of $S_{\lambda_{n\,\text{opt}}}$,

$$2\,n^{-\frac{1}{2}} \sum_{i=1}^{n} \big(Y_{ni} - \bar{Y}_n - x_{ni}^{\mathrm{T}}\,\hat{\beta}_{\text{opt}}\big)\,x_{nij} = \lambda_{nj}\operatorname{sgn}\{(\hat{\beta}_{\text{opt}})_j\}\,.$$

Half the value of the left-hand side is given by

$$\begin{aligned} U_{nj} &\equiv n^{-\frac{1}{2}} \sum_{i=1}^{n} \big\{\epsilon_{ni} - \bar\epsilon_n - x_{ni}^{\rm T}\,(\hat\beta_{\rm opt} - \beta^{n0})\big\}\, x_{nij} \\ &= n^{-\frac{1}{2}} \sum_{i=1}^{n} (\epsilon_{ni} - \bar\epsilon_n) - \bigg\{\bigg(\frac{1}{n}\sum_{i=1}^{n} x_{ni}\, x_{ni}^{\rm T}\bigg)\, n^{\frac{1}{2}}(\hat\beta_{\rm opt} - \beta^{n0})\bigg\}_j \\ &= O_p(1)\,, \end{aligned} \tag{5.17}$$

where the final identity holds for all $j \in \{1,\ldots,q\}$. Hence,

$$P(\mathcal{S} \setminus \widehat{\mathcal{S}}) = \sum_{j=1}^{q} P(j \notin \widehat{\mathcal{S}}) \le \sum_{j=1}^{q} P(\lambda_{nj} \le 2\,|U_{nj}|) \to 0\,, \tag{5.18}$$

where we have used (5.17) and the fact that $\lambda_{nj} \to \infty$ for each $j \in \mathcal{S}$.

5.2. Proof of Theorem 3

Define $\ell_n = \log n$. Recall that the pilot estimator of β is the least-squares estimator $\tilde\beta$, with the property $n^{1/2}\,(\tilde\beta - \beta^0) = \Sigma_n^{-1} Z_n$ where $Z_n = n^{-1/2}\sum_{i\le n} \epsilon_i\, x_i$. Therefore,

$$\hat\epsilon_i = \epsilon_i - \bar\epsilon - x_i^{\rm T}(\tilde\beta - \beta^0)\,. \tag{5.19}$$

Now, $n^{-1}\sum_{i\le n} x_i\, \epsilon_i = n^{-1}\sum_i \epsilon_i\, z_i - \bar\epsilon\, \bar z$, where z_i is as in (3.8). By Kolmogorov's Three-Series Theorem, $(n\,\ell_n^{1+\eta})^{-1/2}\sum_{i\le n} z_i\, \epsilon_i \to 0$ with probability 1, for each $\eta > 0$; see, for example, Petrov (1975, p.274). More simply, by the Law of the Iterated Logarithm, $(n\,\ell_n^{\eta})^{-1/2}\sum_{i\le n} \epsilon_i \to 0$. Hence, for some $\eta > 0$,

$$\bigg(\frac{n}{\ell_n^{1+\eta}}\bigg)^{\frac{1}{2}} (\tilde\beta - \beta^0) \to 0\,, \quad \max_{1\le i\le n} |\hat\epsilon_i - \epsilon_i| = O\bigg\{\bigg(\frac{\ell_n^{1+\eta}}{n}\bigg)^{\frac{1}{2}}\bigg\}\,, \tag{5.20}$$

both results holding with probability 1. Note too that, defining $\mathcal{Z}$ to be the set of data $(X_1, Y_1), \ldots, (X_n, Y_n)$, we have,

$$\hat\sigma^2 \equiv E\big\{(\epsilon_1^*)^2 \,\big|\, \mathcal{Z}\big\} = \frac{1}{n}\sum_{i=1}^{n} \hat\epsilon_i^2\,, \tag{5.21}$$

$$E\big\{(\epsilon_1^*)^2\, I(|\epsilon_1^*| > c) \,\big|\, \mathcal{Z}\big\} = \frac{1}{n}\sum_{i=1}^{n} \hat\epsilon_i^2\, I(|\hat\epsilon_i| > c)\,. \tag{5.22}$$

Assumption (3.8) entails that for a constant $C > 0$, $\|x_i\| \le C$ for all i. This property, (5.19) and the second part of (5.20) imply that, with probability 1, for all sufficiently large n,

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^2\, I(|\hat{\epsilon}_i|>c) \leq \frac{2}{n}\sum_{i=1}^{n}\big\{(\epsilon_i-\bar{\epsilon})^2 + C\,\|\tilde{\beta}-\beta^0\|^2\big\}\, I\big(|\epsilon_i|>\tfrac{1}{2}\,c\big)$$
$$\to 2\,E\big\{\epsilon_1^2\, I\big(|\epsilon_1|>\tfrac{1}{2}\,c\big)\big\}\,. \tag{5.23}$$

Results (5.22) and (5.23) imply that

$$P\bigg[\lim_{c\to 0}\,\limsup_{n\to\infty}\, E\big\{(\epsilon_1^*)^2\, I(|\epsilon_1^*|>c)\,\big|\,\mathcal{Z}\big\}=0\bigg]=1\,. \tag{5.24}$$

Similarly, (5.21) can be used to show that

$$\hat{\sigma}\to\sigma \quad \text{with probability } 1\,. \tag{5.25}$$

It follows from the first part of (5.20) that, provided $m=m(n)\to\infty$ and $m=O(n/\ell_n^{1+\eta})$ for some $\eta>0$,

$$m^{\frac{1}{2}}\max_{1\leq j\leq q}|\tilde{\beta}_j|\to 0\,,\quad m^{\frac{1}{2}}\max_{q+1\leq j\leq p}|\tilde{\beta}_j|\to\infty \tag{5.26}$$

with probability 1. The first three parts of (3.4), and (3.5), are direct consequences of (3.8); the fourth part of (3.4), in the bootstrap world, is equivalent to (5.25); the bootstrap version of (3.6) is implied by (5.24); and the bootstrap version of (3.7) is equivalent to (5.26). Therefore by Theorem 2, with $\tilde{\lambda}$ defined as at (2.5), we have $\tilde{\lambda}_j/m^{1/2}\to\infty$ with probability 1 for $1\leq j\leq q$, and $\tilde{\lambda}_j/m^{1/2}\to 0$ with probability 1 for $q+1\leq j\leq p$. Hence, with $\hat{\lambda}$ as at (2.5),

$$\begin{aligned}
&\frac{\hat{\lambda}_j}{n^{\frac{1}{2}}}\to\infty \quad \text{with probability 1 for} \quad 1\leq j\leq q\,,\\
&\frac{\hat{\lambda}_j}{n^{\frac{1}{2}}}\to 0 \quad \text{with probability 1 for} \quad q+1\leq j\leq p\,.
\end{aligned} \tag{5.27}$$

This is equivalent to part (a) of Theorem 3.

Defining $\Delta=n^{1/2}(\beta-\beta^0)$ and $Z_n=n^{-1/2}\sum_i x_i\,\epsilon_i$, we have

$$\begin{aligned}
S_{\hat{\lambda}}(\beta) = {} & \Delta^{\mathrm{T}}\,\Sigma_n\,\Delta - 2\,\Delta^{\mathrm{T}}\,Z_n + \sum_{j=1}^{q}\Big(\frac{\hat{\lambda}_j}{n^{\frac{1}{2}}}\Big)\,|\Delta_j| \\
& + \sum_{j=q+1}^{p}\Big(\frac{\hat{\lambda}_j}{n^{\frac{1}{2}}}\Big)\,\mathrm{sgn}(\beta_j^0)\,\Delta_j \\
& + \text{terms not depending on } \beta
\end{aligned} \tag{5.28}$$

for all β such that $\max_{q+1\leq j\leq p}|\beta_j-\beta_j^0|\leq\frac{1}{2}\min_{q+1\leq j\leq p}|\beta_j^0|$. Using (5.27) and (5.28), and taking $\beta=\bar{\beta}$ in the vector $\Delta=n^{1/2}(\beta-\beta^0)$, we deduce that for all sufficiently large n, the minimizer, in Δ, of the far right-hand side of (5.28) converges weakly to the random vector $\xi = (\xi_1, \ldots, \xi_p)^{\mathrm{T}}$ that minimizes $\Delta^{\mathrm{T}}\,\Sigma\,\Delta - 2\,\Delta^{\mathrm{T}} Z$, subject to the constraint $\xi_1 = \cdots = \xi_q = 0$, where Z is Normal $\mathrm{N}(0, \sigma^2\,\Sigma)$. This establishes part (c) of Theorem 3.

A proof of part (b) is similar to that of part (b) in Theorem 2. Indeed, it suffices to show that $P(\mathcal{T} \setminus \widehat{\mathcal{T}}) \to 0$. For this we use the fact that, with V_{nj} given by (5.17) with $\hat{\beta}$ replaced by $\bar{\beta}$, we have, in place of (5.18),

$$P(\mathcal{T} \setminus \widehat{\mathcal{T}}) \leq \sum_{j=1}^{q} P\big(n^{-\frac{1}{2}}\hat{\lambda}_j \leq 2\,|V_{nj}|\big) \to 0\,,$$

where the limit result follows from part (a) of Theorem 3 and the fact that, for each $j \in \{1, \ldots, q\}$, $V_{nj} = O_p(1)$.

Acknowledgement

Research of Hall was supported by an Australian Research Council grant. Research of Park was supported by the Korea Science and Engineering Foundation(KOSEF) grant funded by the Korea government(MOST) (No. R01-2007-000-10143-0).

References

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37**, 373-384.

Chen, S. B, Donoho, D. L. and Saunders, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM Rev.* **43**, 129-159.

Donoho, D. L. (2006a). For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Comm. Pure Appl. Math.* **59**, 797–829.

Donoho, D. L. (2006b). For most large underdetermined systems of equations, the minimal ℓ_1-norm near-solution approximates the sparsest near-solution. *Comm. Pure Appl. Math.* **59**, 907-934.

Donoho, D. L. and Elad, M. (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proc. Nat. Acad. Sci. USA* **100**, 2197-2202.

Donoho, D. L. and Huo, X. (2001). Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Theory* **47**, 2845-2862.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: asymptopia? With discussion and a reply by the authors. *J. Roy. Statist. Soc. Ser. B* **57**, 301-369.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Ferris, M. C., Voelker, M. M. and Zhang, H. H. (2004). Model building with likelihood basis pursuit. *Optim. Methods Softw.* **19**, 577-594.

Gao, H.-Y. (1998). Wavelet shrinkage denoising using the non-negative garrote. *J. Comput. Graph. Statist.* **7**, 469-488.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28**, 1356-1378.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34**, 1436-1462.

Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization.* 2nd edition. Springer, New York.

Petrov, V. V. (1975). *Sums of Independent Random Variables.* Springer, Berlin.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc.* B **58**, 267–288.

Tropp, J. A. (2005). Recovery of short, complex linear combinations via ℓ_1 minimization. *IEEE Trans. Inform. Theory* **51**, 1568-1570.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Department of Mathematics and Statistics, The University of Melbourne, VIC 3010, Australia.

E-mail: halpstat@ms.unimelb.edu.au

Department of Statistics, Seoul National University, Seoul 151–747, Korea.

E-mail: silverryuee@gmail.com

Department of Statistics, Seoul National University, Seoul 151–747, Korea.

E-mail: bupark2000@gmail.com

(Received August 2007; accepted November 2007)