Statistics and Its Interface Volume 6 (2013) 231–242

Sparse bridge estimation with a diverging number of parameters

Sunghoon Kwon, Yongdai Kim and Hosik Choi*,†

The Bridge estimator with ℓ_ν^ν-penalty for some $\nu > 0$ is one of the popular choices in penalized linear regression models. It is known that, when $\nu \leq 1$, the Bridge estimator produces sparse models which allow us to control the model complexity. However, when $\nu = 1$, the Bridge estimator fails to identify the correct model since it requires certain strong sufficient conditions that are hard to hold in general, and when $\nu > 1$, it achieves no sparsity in parameter estimation. In this paper, we propose the sparse Bridge estimator that is developed to find the correct sparse version of the Bridge estimator when $\nu \geq 1$. Theoretically, the sparse Bridge estimator is asymptotically equivalent to the oracle Bridge estimator when the number of predictive variables diverges to infinity but less than the sample size. Here, the oracle Bridge estimator is an ideal Bridge estimator obtained by deleting all irrelevant predictive variables in advance. Hence, the sparse Bridge estimator naturally inherits the properties of the Bridge estimator without losing correct model identification asymptotically. Numerical studies show that the sparse Bridge estimator can outperform other penalized estimators with a finite sample.

AMS 2000 subject classifications: Primary 62J05; secondary 62J07.
Keywords and phrases: Bridge, Diverging number of parameters, Lasso, Regression, Ridge, Variable selection.

1. INTRODUCTION

Consider the linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon},$$

where $\mathbf{y} = (y_1, \ldots, y_n)'$ is the response vector, $\mathbf{X} = (X_1, \ldots, X_p)$ is the $n \times p$ design matrix, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ is the random error vector, and $\boldsymbol{\beta}^*$ is the true parameter vector. For given $\nu > 0$, the Bridge estimator [9] is defined as

$$\hat{\boldsymbol{\beta}}^{B,\gamma} = \arg\min_{\boldsymbol{\beta}} \Big\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/2n + \gamma\|\boldsymbol{\beta}\|_\nu^\nu \Big\} \tag{1}$$

for some $\gamma > 0$, where $\|\boldsymbol{\beta}\|_\nu = (\sum_{j=1}^p |\beta_j|^\nu)^{1/\nu}$ is the usual ℓ_ν-norm operator. As a special case, when $\nu = 1$, the Bridge estimator is known as the least absolute shrinkage and selection operator (Lasso) estimator [21],

$$\hat{\boldsymbol{\beta}}^{L,\gamma} = \arg\min_{\boldsymbol{\beta}} \Big\{ \boldsymbol{\beta}'(\mathbf{X}'\mathbf{X})\boldsymbol{\beta}/2n - \mathbf{y}'\mathbf{X}\boldsymbol{\beta}/n + \gamma\|\boldsymbol{\beta}\|_1 \Big\}, \tag{2}$$

that is one of the popular choices for the variable selection problems. Another example, when $\nu = 2$, is the Ridge estimator [10],

$$\hat{\boldsymbol{\beta}}^{R,\gamma} = \arg\min_{\boldsymbol{\beta}} \Big\{ \boldsymbol{\beta}'(\mathbf{X}'\mathbf{X}/2n + \gamma I_{p\times p})\boldsymbol{\beta} - \mathbf{y}'\mathbf{X}\boldsymbol{\beta}/n \Big\}, \tag{3}$$

that properly handles the collinearity problems [10], where the least square estimator is expected to perform worse.

For years, many penalized estimators have been developed for variable selection problems. Especially, some estimators are asymptotically pursuing the equivalent performance to the oracle estimator:

$$\hat{\boldsymbol{\beta}}^{oR} = \arg\min_{\beta_j = 0, j \in \{k:\beta_k^* = 0\}} \Big\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/2n \Big\}. \tag{4}$$

This ideal behavior is described as the oracle property by [7]. Some typical examples are the smoothly clipped absolute deviation (Scad) estimator [7], adaptive Lasso estimator [26], Bridge estimator with $\nu < 1$ [11], and minimax concave penalty (Mcp) estimator [24]. These estimators are practically better in selecting variables compared to the Bridge estimator with $\nu \geq 1$ since they are selection consistent asymptotically.

However, there still exist several examples that present the practical excellence of the Bridge estimators regardless of the oracle property described above. The Lasso estimator in (2) achieves higher prediction accuracy when the true model includes relatively smaller coefficients compared to the sample size [26]. And the Ridge estimator in (3) properly handles the collinearity problems by stabilizing the variance of the least square estimator [4, 10]. Another nice alternative is the elastic net (Enet) estimator [25] defined as

*We are grateful to the anonymous referees and the editor for their helpful comments. Kim's research was supported by the National Research Foundation of Korea grant number 20100012671 funded by the Korea government. Choi's research was supported by Basic Science Research Program through National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0003377).
†Corresponding author.

(5)
$$\hat{\boldsymbol{\beta}}^{eN,\gamma,\lambda} = \arg\min_{\boldsymbol{\beta}}\Big\{\boldsymbol{\beta}'\Big(\frac{\mathbf{X}'\mathbf{X}+\gamma I_{p\times p}}{1+\gamma}\Big)\boldsymbol{\beta} - 2\mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1\Big\}.$$

Although the Enet estimator is not any Bridge estimator, it can be thought of as an improved version of two Bridge estimators, the Lasso and Ridge. Note that, however, these estimators still suffer from the lack of selection consistency that is a major uneasiness for them.

In this paper, we propose the sparse Bridge estimator that is developed to find the correct sparse version of the Bridge estimator defined in (1) when $\nu \geq 1$. The sparse Bridge estimator resolves the major deficiency of the Bridge estimator by imposing sparsity on it without losing its own practical advantage. Theoretically the sparse Bridge estimator is asymptotically equivalent to the oracle Bridge estimator:

(6)
$$\hat{\boldsymbol{\beta}}^{oB,\gamma} = \arg\min_{\beta_j=0, j\in\{k:\beta_k^*=0\}}\Big\{\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2/2n + \gamma\|\boldsymbol{\beta}\|_\nu^\nu\Big\}.$$

As a consequence, the sparse Bridge estimator naturally inherits the properties of the Bridge estimator without losing correct model identification asymptotically.

The rest of the paper is organized as follows. Section 2 introduces the proposed method, and Section 3 proves theoretical properties. Section 4 gives results of various numerical studies. The conclusions and technical details are given in Section 5 and Appendix, respectively.

2. SPARSE BRIDGE ESTIMATION

2.1 Definition and solution

Let $\mathcal{L}(\boldsymbol{\beta}) = \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2/2n$. For a given $\nu \geq 1$, the sparse Bridge estimator is defined as

$$\hat{\boldsymbol{\beta}}^{sB,\gamma,\lambda} = \arg\min_{\boldsymbol{\beta}}\mathcal{Q}^{sB,\gamma,\lambda}(\boldsymbol{\beta}), \tag{7}$$

where

$$\mathcal{Q}^{sB,\gamma,\lambda}(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) + \mathcal{J}^{\gamma,\lambda}(\boldsymbol{\beta}) \tag{8}$$

and $\mathcal{J}^{\gamma,\lambda}(\boldsymbol{\beta}) = \sum_{j=1}^p J^{\gamma,\lambda}(|\beta_j|)$. Here, $J^{\lambda,\gamma}(\cdot)$ is the sparse Bridge penalty that satisfies $J^{\gamma,\lambda}(0) = 0$, and

$$\begin{aligned}\nabla J^{\gamma,\lambda}(t) = {} & (-c^{\gamma,\lambda}t^\nu + \lambda)I[t < a(\lambda-\gamma)] \\ & + \gamma\nu t^{\nu-1}I[t \geq a(\lambda-\gamma)]\end{aligned} \tag{9}$$

for some $a > 0$, where $\nabla J^{\gamma,\lambda}(t) = \partial J^{\gamma,\lambda}(t)/\partial t$ and

$$c^{\gamma,\lambda} = [\lambda - \gamma\nu\{a(\lambda-\gamma)\}^{\nu-1}]/\{a(\lambda-\gamma)\}^\nu$$

for $\lambda \geq \max\{\gamma, \gamma\nu\{a(\lambda-\gamma)\}^{\nu-1}\}$ and $a \geq 1$.

Here are some characteristics on the sparse Bridge penalty which are main motivations of the paper:

- By definition, if $\lambda = \gamma$ then the sparse Bridge penalty is the same as the Bridge penalty.
- And if $\lambda > \gamma$, the sparse Bridge penalty is a smoothly clipped version of the Bridge penalty near the origin.
- The penalty is concave on $(0, a(\lambda-\gamma))$ satisfying $\lim_{t\to 0+}\nabla J^{\gamma,\lambda}(t) = \lambda$ and the same as the Bridge penalty on $[a(\lambda-\gamma), \infty)$ satisfying $J^{\gamma,\lambda}(t) = \gamma\nu t^{\nu-1}$.

Hence the sparse Bridge penalty can select variables that have the same type of shrinkage as the Bridge penalty. And as a referee pointed out, two tuning parameters λ and γ play different roles:

- λ controls the concavity of the sparse Bridge penalty near the origin thresholding small coefficients.
- γ regularizes the amount of shrinkages over the estimated nonzero coefficients by the same way as in the original Bridge.

It is more transparent to see the solutions under orthonormal design, where $\mathbf{X}'\mathbf{X}/n = I$. The solution $\hat{\beta}_j$ satisfies $\nabla J^{\gamma,\lambda}(|\hat{\beta}_j|) + \hat{\beta}_j = \hat{z}_j$, $j = 1,\ldots,p$, where $\hat{z}_j = X_j'\mathbf{y}/n$ is the ordinary least square estimator. Since $\lim_{|\hat{\beta}_j|\to 0}\nabla J^{\gamma,\lambda}(|\hat{\beta}_j|) = \lambda$ and $\nabla J^{\gamma,\lambda}(|\hat{\beta}_j|) = \gamma\nu|\hat{\beta}_j|^{\nu-1}$ for $|\hat{\beta}_j| > a(\lambda-\gamma)$, it is easy to see that small coefficients are excluded by the level of λ, and large solutions are exactly the same as those of the original Bridge estimator with γ. For example, if $\nu = 1$,

$$\hat{\beta}_j^{sB,\gamma,\lambda} = \begin{cases} 0, & |z_j| < \lambda, \\ \frac{a}{a-1}\mathrm{sign}(z_j)(|z_j|-\lambda), & \lambda \leq |z_j| < a(\lambda-\gamma)+\gamma, \\ \mathrm{sign}(z_j)(|z_j|-\gamma), & |z_j| \geq a(\lambda-\gamma)+\gamma, \end{cases}$$

for $\lambda \geq \gamma$, and if $\nu = 2$,

$$\hat{\beta}_j^{sB,\gamma,\lambda} = \begin{cases} 0, & |z_j| < \lambda, \\ u(z_j), & \lambda \leq |z_j| < a(1+2\gamma)(\lambda-\gamma), \\ z_j/(1+2\gamma), & |z_j| \geq a(1+2\gamma)(\lambda-\gamma), \end{cases}$$

for $\lambda \geq \max\{\gamma, \gamma_*\}$ if $\gamma \leq 1/2a$, and $\max\{\gamma,\gamma_*\} < \lambda < \gamma^*$ if $2a\gamma > 1$, where $\gamma_* = a\gamma(1+2\gamma)/(a-1+2a\gamma)$, $\gamma^* = 2a\gamma^2/(2a\gamma-1)$ and $u(z_j)$ is solution of $-c^{\gamma,\lambda}\beta_j^2 + \beta_j + z_j = 0$. See Figure 1 that draws the sparse Bridge penalties and solutions when $\nu = 1$ and 2. See Section 4, for numerical results that show different roles of two tuning parameters.

2.2 Computational algorithm

The sparse Bridge penalty $J^{\gamma,\lambda}(|t|)$ can be decomposed as follows:

$$J^{\gamma,\lambda}(|t|) = \gamma|t|^\nu + \lambda|t| + \tilde{J}^{\gamma,\lambda}(|t|), \tag{10}$$

where $\tilde{J}^{\gamma,\lambda}(|t|)$ is a continuously differentiable concave function. Hence, by letting $\tilde{\mathcal{J}}^{\gamma,\lambda}(\boldsymbol{\beta}) = \mathcal{J}^{\gamma,\lambda}(\boldsymbol{\beta}) - \gamma\|\boldsymbol{\beta}\|_\nu^\nu - \lambda\|\boldsymbol{\beta}\|_1$, we can rewrite (8) by

$$\mathcal{Q}^{sB,\gamma,\lambda}(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) + \gamma\|\boldsymbol{\beta}\|_\nu^\nu + \lambda\|\boldsymbol{\beta}\|_1 + \tilde{\mathcal{J}}^{\gamma,\lambda}(\boldsymbol{\beta})$$

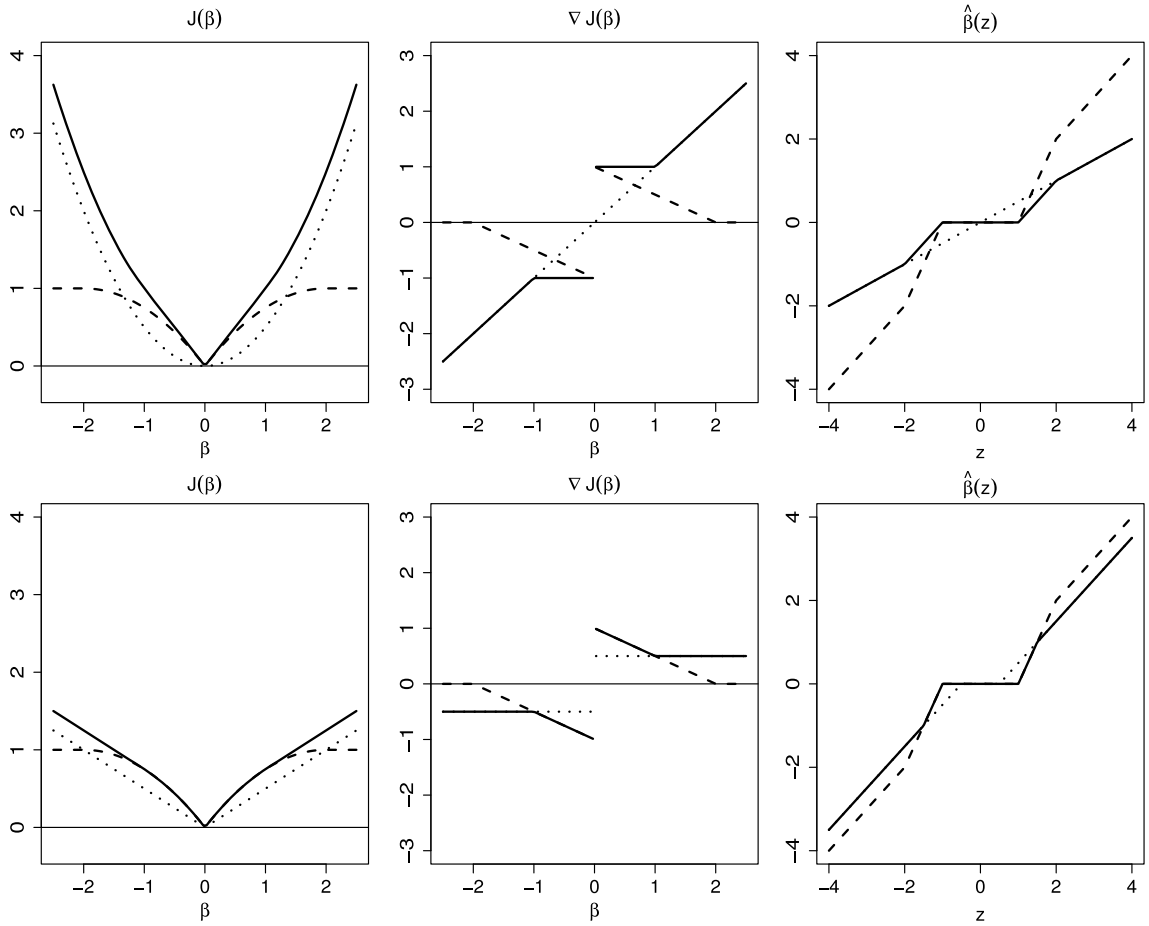

Figure 1. The upper left two panels plot three penalties and derivatives: the Bridge (dotted) with $(\nu, \gamma) = (2, 1/2)$*, Scad (dashed) with* $(\lambda, a) = (1, 2)$ *and sparse Bridge (line) with* $(\nu, \gamma, \lambda, a) = (2, 1/2, 1, 2)$*. The upper right panel shows the corresponding solutions as functions of the ordinary least square estimator* z *under the orthonormal design. Similarly, the below panels draw the plots when* $\nu = 1$*.*

which is a sum of convex and concave functions. Hence, we can find a local minimizer of (8) by use of the convex-concave procedure (CCCP) [23]. Note that the CCCP algorithm is one of powerful optimization algorithms for nonconvex problems [3, 13, 20]. One of the main properties of the CCCP algorithm is that it always converges to a local minimizer [1, 2].

To be specific, let $\nabla \tilde{\mathcal{J}}^{\gamma,\lambda}(\boldsymbol{\beta}) = \partial \tilde{\mathcal{J}}^{\gamma,\lambda}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$. For a given current solution $\hat{\boldsymbol{\beta}}^c$, we update it by minimizing the upper tight convex function,

$$\mathcal{U}^{sB,\gamma,\lambda}(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) + \gamma\|\boldsymbol{\beta}\|_\nu^\nu + \nabla \tilde{\mathcal{J}}^{\gamma,\lambda}(\hat{\boldsymbol{\beta}}^c)'\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1,$$

until it converges. Many efficient optimization algorithms are available to minimize this convex function, since it is simply a ℓ_1-penalized convex problem. For special cases, if $\nu = 1$, then we can use the least angle regression algorithm developed by [6]. For other cases, the predictor-corrector algorithm introduced by [17] can be applied.

3. THEORETICAL PROPERTIES

In this subsection, we prove asymptotic equivalence between the sparse Bridge estimator and oracle Bridge estimator so that it is asymptotically as efficient as the oracle Bridge estimator in (6). This relationship is the same as that of the Scad estimator and oracle estimator in (4) [8, 13].

Without loss of generality, we assume that

$$\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{*\prime}, \boldsymbol{\beta}_2^{*\prime})',$$

where $\boldsymbol{\beta}_1^*$ is a $q \times 1$ vector whose elements are all nonzero, and $\boldsymbol{\beta}_2^*$ is a $(p-q) \times 1$ vector whose elements are all zero. Further we assume that $\varepsilon_i, 1 \le i \le n$ are independent and identically distributed random variables with $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma_0^2$ for some $\sigma_0 < \infty$. Let $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, where $\mathbf{X}_1 = (X_1, \ldots, X_q)$, $\mathbf{X}_2 = (X_{q+1}, \ldots, X_p)$ with $X_j = (x_{1j}, \ldots, x_{nj})', 1 \le j \le p$.

Denote ρ_n and τ_n as the smallest and largest eigenvalues of the matrix $\mathbf{X}'\mathbf{X}/n$, respectively. To allow the number of

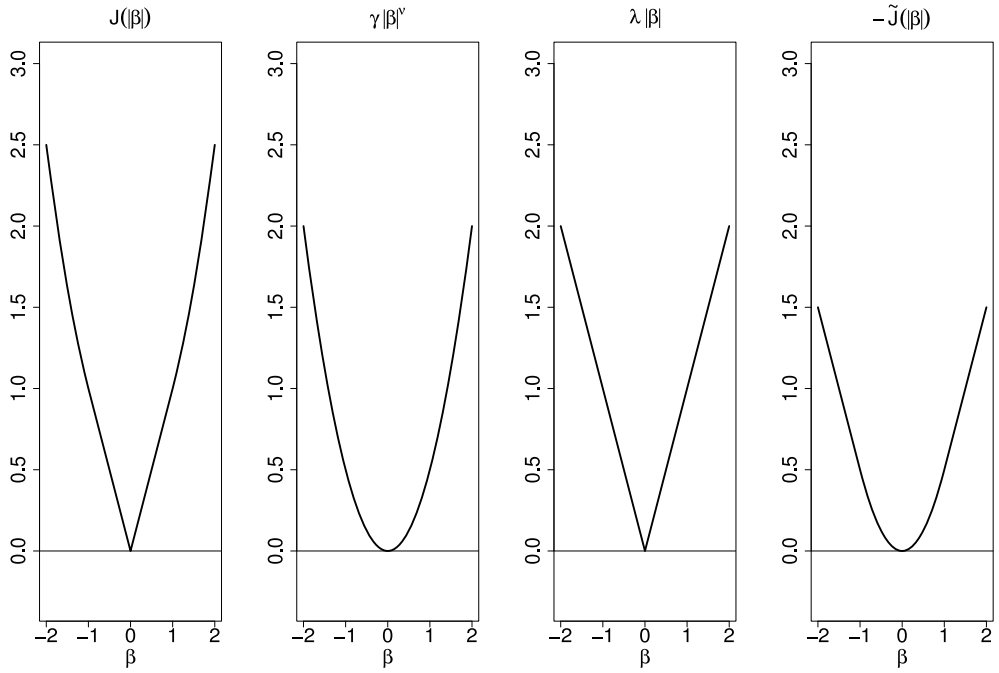

Figure 2. Decomposition of $J^{\gamma,\lambda}(|\boldsymbol{\beta}|) = \gamma|\boldsymbol{\beta}|^{\nu} + \lambda|\boldsymbol{\beta}| - \tilde{J}^{\gamma,\lambda}(|\boldsymbol{\beta}|)$.

parameters $p < n$ to diverge to infinity, we require following regularity conditions with positive constants $\rho_0, \tau_0, b_0, \delta_0, d_0$ and $c_0 < 1$:

(*A*1) $\rho_0 < \rho_n \le \tau_n \le \tau_0$ for all $n \ge 1$.
(*A*2) $\max_{1\le j\le p}\|X_j\|_2^2/n \le b_0$ for all $n \ge 1$.
(*A*3) $\log p/\log n \to c_0$ as $n \to \infty$.
(*A*4) $\max_{1\le j\le q}|\beta_j^*| < d_0$ for all $n \ge 1$.
(*A*5) $\min_{1\le j\le q}|\beta_j^*|/(q/n)^{1/2} \to 0$ as $n \to \infty$.
(*A*6) $\max_{1\le i\le n}\sum_{j=1}^{p} x_{ij}^2/n \to 0$ as $n \to \infty$.
(*A*7) $E|\varepsilon_1|^{2+\delta_0} < \infty$.

Condition (*A*1) is standard for the asymptotic study on linear regression model, and (*A*2) is directly satisfied if we standardize the covariates. Condition (*A*3) specifies the order of p. As pointed out by [27], it deletes the limitation of $p = o(n^{1/4})$ in [8] for the linear regression model. Condition (*A*4) is technically considered for ease of exposition that holds implicitly with fixed p and q. Conditions (*A*5), (*A*6) and (*A*7) are used only to construct the limiting distribution of the oracle Bridge estimator. Note that (*A*5) allows the true coefficients to decrease toward zero.

3.1 Asymptotic properties of the oracle Bridge estimator

First, we give an expected ℓ_2-risk bound of the oracle Bridge estimator.

Theorem 3.1. *Assume that* (*A*1) *holds then*

$$E\|\hat{\boldsymbol{\beta}}_1^{oB,\gamma} - \boldsymbol{\beta}_1^*\|_2 \le \Delta(\boldsymbol{\beta}_1^*, \gamma)/\rho_0 + \sigma_0(q/n\rho_0)^{1/2}, \tag{11}$$

where

$$\Delta(\boldsymbol{\beta}_1^*, \gamma) = \nu\gamma^{1/\nu} q^{\{(2-\nu)/2\nu\}I(\nu<2)}(\gamma\|\boldsymbol{\beta}_1^*\|_\nu^\nu + \sigma_0^2 q/2n)^{(\nu-1)/\nu}.$$

Theorem 3.1 is non-asymptotic and useful for developing asymptotic theories. In the theorem, the second term $\sigma_0(q/n\rho_0)^{1/2}$ is due to the random error that is $O((q/n)^{1/2})$, whereas, the first term yields extra bias that depends on the tuning parameter γ and the size of the true nonzero parameters.

Remark 1. From Theorem 3.1, it is easy to see that the oracle Bridge estimator is consistent if $\Delta(\boldsymbol{\beta}_1^*, \gamma) \to 0$ that is equivalent to

$$\gamma q^{1/2+\{(\nu-2)/2\nu\}I(\nu\ge 2)} \to 0 \tag{12}$$

as $n \to \infty$. Hence, we require smaller γ for larger ν and p. For example, when $\nu = 1$ and $\nu = 2$, $q^{1/2}\gamma \to 0$ implies the consistency of the oracle Lasso and Ridge estimators, respectively.

Let $\nabla_1(\boldsymbol{\beta}) = \partial(\|\boldsymbol{\beta}\|_\nu^\nu)/\partial\boldsymbol{\beta}$ and $\mathbf{X}_1'\mathbf{X}_1/n = \boldsymbol{\Sigma}_1$. The next theorem gives $(q/n)^{1/2}$-consistency and asymptotic normality of the oracle Bridge estimator.

Theorem 3.2. *Assume that* (*A*1)–(*A*7) *hold then*

(a) $\|\hat{\boldsymbol{\beta}}_1^{oB,\gamma} - \boldsymbol{\beta}_1^*\|_2 = O_p((q/n)^{1/2})$,

(b) *for any* $\boldsymbol{\alpha} \in \mathbb{R}^q$ *with* $\|\boldsymbol{\alpha}\|_2 = 1$,

$$n^{1/2}\boldsymbol{\alpha}'\boldsymbol{\Sigma}_1^{1/2}\{\hat{\boldsymbol{\beta}}_1^{oB,\gamma} - \boldsymbol{\beta}_1^* + \gamma\boldsymbol{\Sigma}_1^{-1}\nabla_1(\boldsymbol{\beta}_1^*)\} \to N(0, \sigma_0^2)$$

in distribution, provided

$$\gamma n^{1/2} q^{\{(\nu-2)/2\nu\}I(\nu\geq 2)} \to 0 \tag{13}$$

as $n \to \infty$.

In Theorem 3.2, γ must vanish toward zero faster than the rate of consistency in (12) in Remark 1 by a factor of $(q/n)^{1/2}$ exactly, and this result is consistent to that of [15] with fixed p and q. From Theorem 3.2, we can see that the bias of the Lasso, when $\nu = 1$, depends on the tuning parameter γ only while the bias depends on the size of the true parameters also when $\nu > 1$. Note that Theorem 3.1 plays a major role for the results of Theorem 3.2 since $(q/n)^{1/2}$-consistency directly comes from (11).

3.2 Asymptotic equivalence between the oracle Bridge estimator and sparse Bridge estimator

Next, we prove the optimality of the oracle Bridge estimator for the objective function defined in (8). That is, we provide sufficient conditions under which the sparse Bridge estimator is exactly the same as the oracle Bridge estimator asymptotically.

Let $\Omega^{sB,\gamma,\lambda}$ be the set of all local minimizers of (8). The following theorem is our main result that proves the sparse Bridge estimator is exactly the same as the oracle Bridge estimator asymptotically.

Theorem 3.3. *Assume that* (A1)–(A4) *hold. If* $\lambda \to 0$, $(n/p)^{1/2}\lambda \to \infty$, *and* $\min_{1\leq j\leq q}|\beta_j^*|/\lambda \to \infty$ *then*

(a) (Local optimality) $\mathbf{P}\big(\hat{\boldsymbol{\beta}}^{oB,\gamma} \in \Omega^{sB,\gamma,\lambda}\big) \to 1$,
(b) (Global optimality)

$$\mathbf{P}\big(\hat{\boldsymbol{\beta}}^{oB,\gamma} = \arg\min_{\boldsymbol{\beta}} \mathcal{Q}^{sB,\gamma,\lambda}(\boldsymbol{\beta})\big) \to 1,$$

provided $\gamma \to 0$ *and*

$$(\gamma/\lambda)q^{1/2+\{(\nu-2)/2\nu\}I(\nu\geq 2)} \to 0 \tag{14}$$

as $n \to \infty$.

In Theorem 3.3, the conditions imposed on λ are the same as those of [8]. As [8] pointed out, these conditions show the least favorable rate of $\min_{1\leq j\leq q}|\beta_j^*|$ for oracle variable selection. Note that, for the results to hold, γ must decrease faster than both λ and the rate (12) in Remark 1, but slower than (13). This implies that (14) is stronger than (12) but weaker than (13).

4. NUMERICAL STUDIES

4.1 Simulation studies

In this section, we present simulation studies to evaluate the finite sample performance of the sparse Bridge estimator. Under the linear regression model,

$$y = x'\boldsymbol{\beta}^* + \varepsilon, \quad \varepsilon \sim N(0, \sigma_0^2). \tag{15}$$

The covariate vector $x = (x_1, \ldots, x_p)'$ is set to be a multivariate Gaussian random vector with mean zero and the covariance of x_j and x_k to be $\rho_0^{|j-k|}, j, k = 1, \ldots, p$. We select the true nonzero elements of $\boldsymbol{\beta}^*$ as a sequence that satisfies $\beta_j^* = 1.5 - (j-1)/(q-1), j = 1, \ldots, q$ and $\beta_j^* = 0, j = q+1, \ldots, p-q$. For sample sizes $n = 30, 60$ and 120, we try $p = [n/3]$ and $q = [p/3]$. Here, $[x]$ indicates the smallest integer greater than x.

We investigate two features: (a) prediction error based on independent test data set of size $2n$; (b) number of coefficients that are selected correctly and incorrectly. From 400 independent repetitions, we measure the medians of averaged prediction errors, and the frequency of correctly and incorrectly selected nonzero coefficients. For comparison, we consider six estimators: Lasso, Scad, Ridge, Enet, sparse Bridge with $\nu = 1$ (sLasso), and $\nu = 2$ (sRidge). We set $a = 3.7$ for the Scad, sLasso and sBridge, and choose other tuning parameters using independent validation data of size $n/2$.

Table 1 summarizes the results, and we present some observations:

1. The two sparse Bridge estimators sLasso and sRidge outperform the other methods improving the Lasso and Ridge respectively for almost all cases.
2. When the sample size n is large or the noise level σ_0 is small, the selection consistent estimators (Scad, sLasso and sBridge) perform better than the others (Lasso, Ridge and Enet).
3. When the sample size is small the characteristics of penalties have positive effects on the prediction accuracy, and this shows why the Scad performs worst when $\sigma_0 = 3$ even if it has the oracle property.
4. When $\rho_0 = 0.9$ and $\sigma_0 = 3$, the Ridge performs better than other estimators except the sRidge, which clearly shows that both sparsity and type of shrinkage closely are related to the prediction accuracy with finite sample size. This is similar to the relation between the Enet and the Lasso.
5. The Scad produces most sparse solutions, and the sLasso and sRidge are sparser than the Lasso and Enet respectively. Except the Ridge which cannot control the selectivity, the Lasso has the largest final model, and the order of model size is Lasso>Enet>sRidge>sLasso>Scad. An interesting point is that sRidge is sparser than Enet. This means that the Enet estimator cannot control the sparsity

Table 1. Simulation results comparing the medians of prediction errors, means of correctly and incorrectly selected nonzero coefficients

Prediction error

σ_0	r	n	p	q	Lasso	sLasso	Ridge	sRidge	Scad	Enet
1	0.5	30	10	3	0.1946	**0.1923**	0.3399	**0.1848**	0.2146	0.2001
		60	20	7	0.2220	**0.1682**	0.3593	**0.1636**	0.1817	0.2202
		120	40	13	0.2076	0.1373	0.3888	**0.1289**	**0.1311**	0.1931
	0.9	30	10	3	0.2457	**0.1843**	0.2173	**0.1526**	0.2506	0.2430
		60	20	7	0.2233	**0.1883**	0.2588	**0.1583**	0.2003	0.2197
		120	40	13	0.1883	**0.1593**	0.2341	**0.1285**	0.1603	0.1774
3	0.5	30	10	3	**1.3319**	**1.3664**	1.6174	1.3974	1.7645	1.4234
		60	20	7	1.7928	**1.7438**	1.9212	**1.7542**	2.5425	1.7725
		120	40	13	1.8318	**1.7239**	2.2429	**1.5121**	2.4645	1.7937
	0.9	30	10	3	1.1938	1.0817	**0.8902**	**1.0463**	1.2718	1.3058
		60	20	7	1.5983	1.3026	**1.2839**	**1.1114**	1.7236	1.5402
		120	40	13	1.3841	1.2596	**1.1983**	**0.8158**	1.6262	1.3127

Number of nonzero coefficients correctly estimated

σ_0	r	n	p	q	Lasso	sLasso	Ridge	sRidge	Scad	Enet
1	0.5	30	10	3	2.92	2.74	3.00	2.78	2.66	2.92
		60	20	7	6.98	6.88	7.00	6.90	6.81	6.97
		120	40	13	13.00	12.98	13.00	12.98	12.96	13.00
	0.9	30	10	3	2.51	2.24	3.00	2.43	2.04	2.49
		60	20	7	6.70	6.24	7.00	6.48	5.99	6.71
		120	40	13	12.89	12.53	13.00	12.65	12.23	12.88
3	0.5	30	10	3	2.33	2.07	3.00	2.12	1.91	2.32
		60	20	7	6.23	5.51	7.00	5.54	4.82	6.23
		120	40	13	12.46	11.59	13.00	11.64	10.47	12.45
	0.9	30	10	3	1.74	1.53	3.00	1.72	1.46	1.73
		60	20	7	4.91	3.97	7.00	4.71	3.53	4.87
		120	40	13	10.64	9.08	13.00	10.52	7.80	10.61

Number of nonzero coefficients incorrectly estimated

σ_0	r	n	p	q	Lasso	sLasso	Ridge	sRidge	Scad	Enet
1	0.5	30	10	3	2.80	1.03	7.00	1.30	1.01	2.84
		60	20	7	4.97	1.18	13.00	1.42	1.11	4.88
		120	40	13	9.18	1.30	27.00	1.61	1.26	8.77
	0.9	30	10	3	2.44	0.80	7.00	1.07	0.73	2.26
		60	20	7	3.73	0.96	13.00	1.38	0.98	3.65
		120	40	13	5.78	0.97	27.00	1.58	0.87	5.66
3	0.5	30	10	3	2.75	1.70	7.00	2.09	1.71	2.65
		60	20	7	4.79	2.33	13.00	3.03	2.31	4.73
		120	40	13	8.80	4.02	27.00	4.63	4.00	8.62
	0.9	30	10	3	1.88	1.08	7.00	1.29	1.11	1.81
		60	20	7	3.21	1.12	13.00	1.93	1.42	3.13
		120	40	13	5.64	1.84	27.00	2.66	2.27	5.40

properly, and this is partly explained by the selection consistency of the sRidge.

To sum up, we can see that the sLasso and sRidge can be the correct sparse version of the Lasso and Ridge respectively, and the sLasso and sRidge can be alternatives to other penalized estimators. We note that when the sample size is sufficiently large, one can expect the Scad performs best as the asymptotic property shows [8, 14]. However when the sample size is small we may carefully choose an appropriate penalty. We recommend that one may use sLasso when the sample size is small and the noise level is large, and sRidge when there exists the collinearity problems as the results of simulation studies show.

Table 2. Performance results from 100 random partitions of the data (standard errors)

p		Lasso	sLasso	Ridge	sRidge	Scad	Enet
20	Prediction error	0.4259	0.4196	**0.3671**	0.3842	0.4493	0.4070
		(0.0181)	(0.0147)	(0.0135)	(0.0103)	(0.0135)	(0.0126)
	No. of selected variables	8.97	4.98	20	6.84	2.79	8.70
		(0.19)	(0.17)	(0)	(0.20)	(0.06)	(0.18)
30	Prediction error	0.4319	0.4254	**0.3891**	0.3979	0.4556	0.4150
		(0.0186)	(0.0158)	(0.0151)	(0.0122)	(0.0145)	(0.0136)
	No. of selected variables	9.44	5.10	30	8.13	2.83	9.07
		(0.19)	(0.17)	(0)	(0.23)	(0.05)	(0.19)
40	Prediction error	0.4291	0.4206	0.3910	**0.3901**	0.4645	0.4125
		(0.0184)	(0.0155)	(0.0158)	(0.0108)	(0.0145)	(0.0134)
	No. of selected variables	10.58	5.67	40	9.23	2.98	10.30
		(0.22)	(0.19)	(0)	(0.50)	(0.05)	(0.21)

4.2 Real data example

4.2.1 Gene TRIM32

We employ the data set used in [19], which consists of gene expression levels of 18,975 genes obtained from 120 rats. The main objective of the analysis is to find genes that are correlated with gene TRIM32 known to cause Bardet-Biedl syndrome. As was done by [12], we first select 3,000 genes with the largest variance in expression level, and then choose the top $p = 20, 30$ and 40 genes that have the largest absolute correlation with gene TRIM32 among the selected 3,000 genes.

We compare prediction accuracy and number of variables selected in the model for the methods used in the previous subsection. The data set is randomly divided into three parts: 70 samples for training, 30 samples for validating and the other samples for testing.

Table 2 shows the results: all values are arithmetic means of 100 replicated experiments. It is easy to see that the Ridge has the smallest prediction errors and the sRidge follows it showing almost the same prediction errors but using less than 10 variables. Given the way of the prescreening method, the good performance of the Ridge is intuitive in part since there are strong correlations among predictive variables, and this can be a reason why the Enet shows smaller prediction errors also.

On the other hand, the Scad has the largest prediction errors including only 3 variables, which seems to be too sparse. Further, the difference between the Scad and the others is relatively large. This shows that there are cases where we need to choose penalties carefully, and the sRidge carries out this goal well in this example.

4.2.2 Diabetes study

As aforementioned above, two tuning parameters play different roles: one for selection and the other for shrinkage. To address this issue numerically, we additionally investigate the trajectory of nonzero coefficients of diabetes study data [6]. This data includes 442 samples measured on 10 baseline variables (age, sex, bmi, map and 6 serum measurements) from diabetes patients. The response of interest is a quantitative measure of disease progression one year after baseline.

Figure 3 shows the number of estimated nonzero coefficients (selection) and their ℓ_1-norm (shrinkage) for sLasso ($\nu = 1$) and sRidge ($\nu = 2$). In the plot, we fix four values of λ, and for each λ we find solution paths over 10 values of $\gamma = \lambda/2k, k = 1, \ldots, 10$. It is easy to see that the model size increases as λ decreases, however, it rarely changes over γ. On the other hand, the estimated ℓ_1-norm increases as γ decreases. See Figure 4 that draws the paths of estimated coefficients. This shows that the two tuning parameters λ and γ play clearly as they are expected.

5. CONCLUDING REMARK

We proposed the sparse Bridge estimator to improve the usual Bridge estimator, and proved that it is selection consistent under mild conditions. We confirmed the theoretical results via numerical studies which show how the sparse Bridge estimator behaves and produces better prediction accuracy and variable selectivity than the original one.

Besides ν, the sparse Bridge penalty has two more tuning parameters λ and γ and this may cause computational burdens in practice compared with the Lasso and Scad. We refer to two simpler ways of choosing the tuning parameters. One is to use the universal penalty level for λ fixed with $\lambda = \sigma_0\{(2/n)\log p\}^{1/2}$ as was done by [5, 24], and the other one is to use a heuristic choice of γ fixed with $\gamma = \hat{\gamma}$, where $\hat{\gamma}$ is the optimal one for the original Bridge penalty. Although we do not present the results here, such a choice of γ performs quite well. Once a tuning parameter is fixed, one can apply known selection methods. For example, the cross-validation method, training-validating-testing procedure, AIC, BIC, and GCV, where the generalized degrees of freedom is defined similar to [7, 21] and [22].

In this paper, we only consider linear regression models with diverging $p < n$, however we believe that the proposed method can be extended to quite general models such

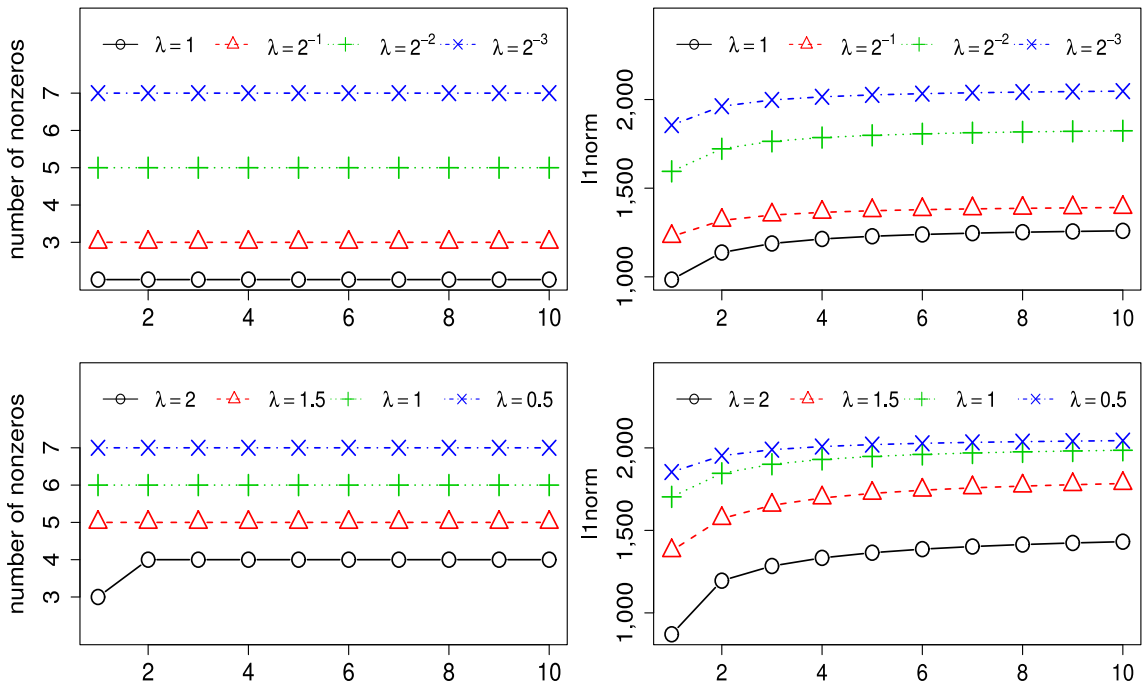

Figure 3. Estimated number of nonzero coefficients and their ℓ_1-norm over γ for fixed λ: sLasso($\nu = 1$, upper) and sRidge($\nu = 2$, below).

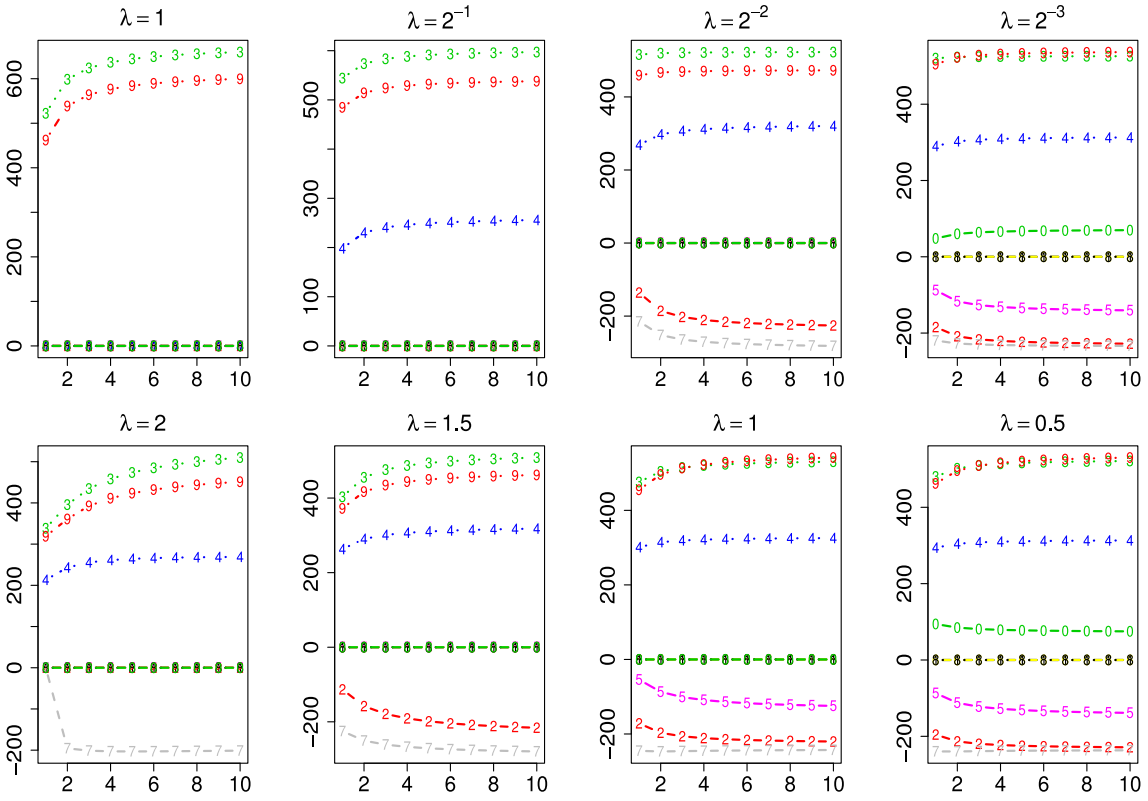

Figure 4. Estimated coefficient path over γ for fixed λ: sLasso($\nu = 1$, upper) and sRidge($\nu = 2$, below).

as generalized linear models or maximum likelihood estimations with $p > n$ [13, 16, 24]. Another challenging problem is to develop uniqueness conditions as [14] that enhances the applicability of the sparse Bridge penalty. We leave these problems for future work.

APPENDIX

For convenience, we define $\mathbf{S}(\boldsymbol{\beta}) = (\mathbf{S}_1(\boldsymbol{\beta})', \mathbf{S}_2(\boldsymbol{\beta})')'$, where $\mathbf{S}_1(\boldsymbol{\beta}) = (S_1(\boldsymbol{\beta}), \ldots, S_q(\boldsymbol{\beta}))'$ and $\mathbf{S}_2(\boldsymbol{\beta}) = (S_{q+1}(\boldsymbol{\beta}), \ldots, S_p(\boldsymbol{\beta}))'$ with $S_j(\boldsymbol{\beta}) = X_j'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})/n, 1 \le j \le p$.

Proof of Theorem 3.1. From direct calculation,

$$\hat{\boldsymbol{\beta}}_1^{oB,\gamma} - \boldsymbol{\beta}_1^* = \boldsymbol{\Sigma}_1^{-1}\{ - \mathbf{S}_1(\hat{\boldsymbol{\beta}}^{oB,\gamma}) + \mathbf{X}_1'\boldsymbol{\varepsilon}/n\}. \tag{16}$$

From $(A1)$,

$$\begin{aligned} E\|\boldsymbol{\Sigma}_1^{-1}\mathbf{X}_1'\boldsymbol{\varepsilon}/n\|_2^2 &= \operatorname{trace}\{E(\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1/n)^{-2}\mathbf{X}_1'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'/n^2)\} \\ &\le q\sigma_0^2/(n\rho_0), \end{aligned}$$

hence, it follows that

$$E\|\hat{\boldsymbol{\beta}}_1^{oB,\gamma} - \boldsymbol{\beta}_1^*\|_2 \le E\|\mathbf{S}_1(\hat{\boldsymbol{\beta}}^{oB,\gamma})\|_2/\rho_0 + \sigma_0(q/n\rho_0)^{1/2}. \tag{17}$$

If $\nu = 1$, by the Karush-Kuhn-Tucker (KKT) optimality conditions (see, [18]), $\hat{\boldsymbol{\beta}}_1^{oB,\gamma}$ satisfies

$$\begin{cases} S_j(\hat{\boldsymbol{\beta}}^{oB,\gamma}) = \gamma\operatorname{sign}(\hat{\beta}_j^{oB,\gamma}), & \hat{\beta}_j^{oB,\gamma} \ne 0, \\ |S_j(\hat{\boldsymbol{\beta}}^{oB,\gamma})| \le \gamma, & \hat{\beta}_j^{oB,\gamma} = 0, \end{cases} \tag{18}$$

for $1 \le j \le q$, and if $\nu > 1$, it satisfies

$$S_j(\hat{\boldsymbol{\beta}}^{oB,\gamma}) = \gamma\nu|\hat{\beta}_j^{oB,\gamma}|^{\nu-1}\operatorname{sign}(\hat{\beta}_j^{oB,\gamma}) \tag{19}$$

for $1 \le j \le q$. Hence,

$$\|\mathbf{S}_1(\hat{\boldsymbol{\beta}}^{oB,\gamma})\|_2 \le \gamma\nu\|\hat{\boldsymbol{\beta}}^{oB,\gamma}\|_{2(\nu-1)}^{\nu-1}. \tag{20}$$

On the other hand, by the definition of $\hat{\boldsymbol{\beta}}_1^{oB,\gamma}$,

$$\begin{aligned} &\gamma\|\hat{\boldsymbol{\beta}}_1^{oB,\gamma}\|_\nu^\nu \\ &\quad\le \gamma\|\boldsymbol{\beta}_1^*\|_\nu^\nu + \|\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1^*\|_2^2/2n - \|\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1^{oB,\gamma}\|_2^2/2n \\ &\quad\le \gamma\|\boldsymbol{\beta}_1^*\|_\nu^\nu + \|\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1^*\|_2^2/2n - \|\mathbf{y} - \Pi_1\mathbf{y}\|_2^2/2n, \end{aligned}$$

where $\Pi_1 = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$. Hence,

$$\gamma\|\hat{\boldsymbol{\beta}}_1^{oB,\gamma}\|_\nu^\nu \le \gamma\|\boldsymbol{\beta}_1^*\|_\nu^\nu + \|\Pi_1\boldsymbol{\varepsilon}\|_2^2/2n. \tag{21}$$

Now, if $\nu < 2$, from (20) and (21), it follows that

$$\begin{aligned} \|\mathbf{S}_1(\hat{\boldsymbol{\beta}}^{oB,\gamma})\|_2 &\le \nu\gamma q^{(2-\nu)/2\nu}\|\hat{\boldsymbol{\beta}}_1^{oB,\gamma}\|_\nu^{\nu-1} \\ &\le \nu\gamma^{1/\nu}q^{(2-\nu)/2\nu}\big(\gamma\|\boldsymbol{\beta}_1^*\|_\nu^\nu + q\sigma_0^2/2n\big)^{(\nu-1)/\nu} \end{aligned} \tag{22}$$

since $E\big(\|\Pi_1\boldsymbol{\varepsilon}\|_2^2/2n\big) = q\sigma_0^2/2n$. From (17) and (22),

$$E\|\hat{\boldsymbol{\beta}}_1^{oB,\gamma} - \boldsymbol{\beta}_1^*\|_2 \le \Delta(\boldsymbol{\beta}_1^*, \gamma)/\rho_0 + \sigma_0(q/n\rho_0)^{1/2}.$$

Similarly, if $\nu \ge 2$, from (20) and (21),

$$\begin{aligned} \|\mathbf{S}_1(\hat{\boldsymbol{\beta}}^{oB,\gamma})\|_2 &\le \nu\gamma\|\hat{\boldsymbol{\beta}}_1^{oB,\gamma}\|_\nu^{\nu-1} \\ &\le \nu\gamma^{1/\nu}\big(\gamma\|\boldsymbol{\beta}_1^*\|_\nu^\nu + \|\Pi_1\boldsymbol{\varepsilon}\|_2^2/2n\big)^{(\nu-1)/\nu}. \end{aligned}$$

Hence, from (17) again,

$$E\|\hat{\boldsymbol{\beta}}_1^{oB,\gamma} - \boldsymbol{\beta}_1^*\|_2 \le \Delta(\boldsymbol{\beta}_1^*, \gamma)/\rho_0 + \sigma_0(q/n\rho_0)^{1/2}.$$

This completes the proof. □

Proof of Theorem 3.2. It is easy to see that $(A4)$ and (13) imply

$$\Delta(\boldsymbol{\beta}_1^*, \gamma) = O_p((q/n)^{1/2}).$$

Hence, from Theorem 3.1, it follows that

$$\|\hat{\boldsymbol{\beta}}_1^{oB,\gamma} - \boldsymbol{\beta}_1^*\|_2 = O_p((q/n)^{1/2}). \tag{23}$$

Next, we will show (b). From (16) and (19), we have

$$\begin{aligned} &n^{1/2}\boldsymbol{\alpha}'\boldsymbol{\Sigma}_1^{1/2}\{\hat{\boldsymbol{\beta}}_1^{oB,\gamma} - \boldsymbol{\beta}_1^* + \boldsymbol{\Sigma}_1^{-1}\gamma\nabla_1(\boldsymbol{\beta}_1^*)\} \\ &\quad= \boldsymbol{\alpha}'\boldsymbol{\Sigma}_1^{-1/2}\mathbf{X}_1'\boldsymbol{\varepsilon}/n^{1/2} \\ &\qquad+ n^{1/2}\boldsymbol{\alpha}'\boldsymbol{\Sigma}_1^{-1/2}\gamma\{\nabla_1(\boldsymbol{\beta}_1^*) - \nabla_1(\hat{\boldsymbol{\beta}}_1^{oB,\gamma})\}. \end{aligned}$$

It is a standard [11] that, under $(A1)$–$(A7)$,

$$\boldsymbol{\alpha}'\boldsymbol{\Sigma}_1^{-1/2}\mathbf{X}_1'\boldsymbol{\varepsilon}/n^{1/2} \to N(0, \sigma_0^2)$$

in distribution for any $\boldsymbol{\alpha} \in \mathbb{R}^q$ with $\|\boldsymbol{\alpha}\|_2 = 1$. Hence, it suffices to show that

$$\Delta = n^{1/2}\boldsymbol{\alpha}'\boldsymbol{\Sigma}_1^{-1/2}\gamma\{\nabla_1(\boldsymbol{\beta}_1^*) - \nabla_1(\hat{\boldsymbol{\beta}}_1^{oB,\gamma})\} = o_p(1).$$

If $\nu = 1$, from (A5) and (23), $\operatorname{sign}(\hat{\beta}_j^{oB,\gamma}) = \operatorname{sign}(\beta_j^*)$ for all $1 \le j \le p$ and sufficiently large n, which implies

$$|\Delta|^2 \le \gamma^2 n\|\nabla_1(\boldsymbol{\beta}_1^*) - \nabla_1(\hat{\boldsymbol{\beta}}_1^{oB,\gamma})\|_2^2/\rho_0 = 0.$$

If $\nu > 1$, by Taylor's expansion, there exists a $\tilde{\boldsymbol{\beta}}_1$ that lies between $\hat{\boldsymbol{\beta}}_1^{oB,\gamma}$ and $\boldsymbol{\beta}_1^*$ such that

$$\begin{aligned} \|\nabla_1(\boldsymbol{\beta}_1^*) - \nabla_1(\hat{\boldsymbol{\beta}}_1^{oB,\gamma})\|_2^2 &= \|\nabla_2(\tilde{\boldsymbol{\beta}}_1)(\hat{\boldsymbol{\beta}}_1^{oB,\gamma} - \boldsymbol{\beta}_1^*)\|_2^2 \\ &\le \nu^2(\nu-1)^2\max_{1\le j\le q}|\tilde{\beta}_j|^{2(\nu-2)}\|(\hat{\boldsymbol{\beta}}_1^{oB,\gamma} - \boldsymbol{\beta}_1^*)\|_2^2. \end{aligned}$$

If $1 < \nu < 2$, $(A5)$ and (23) imply

$$\max_{1\le j\le q}|\tilde{\beta}_j|^{2(\nu-2)} = \min_{1\le j\le q}\{O_p(|\beta_j^*|)\}^{2(\nu-2)} = o_p((q/n)^{\nu-2}).$$

Hence, from (13), we have

$$|\Delta|^2 \le \gamma^2 n o_p((q/n)^{\nu-2}) O_p(q/n) = o_p(1).$$

Similarly, if $\nu \ge 2$, from $(A4)$, (23), and (13),

$$|\Delta|^2 \le \gamma^2 n O_p(1) O_p(q/n) = o_p(1).$$

This completes the proof. □

We need two lemmas to prove Theorem 3.3.

Lemma 1. *For given $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$, if $\hat{\boldsymbol{\beta}}$ satisfies*

$$\begin{cases} S_j(\hat{\boldsymbol{\beta}}) = \gamma\nu|\hat{\beta}_j|^{\nu-1}\text{sign}(\hat{\beta}_j), & |\hat{\beta}_j| > a(\lambda-\gamma), \\ |S_j(\hat{\boldsymbol{\beta}})| < \lambda, & \hat{\beta}_j = 0, \end{cases}$$

for $1 \le j \le p$, then $\hat{\boldsymbol{\beta}} \in \Omega^{sB,\gamma,\lambda}$.

Proof. It suffices to show that there exists a $\delta > 0$ such that $\mathcal{Q}^{sB,\gamma,\lambda}(\boldsymbol{\beta}) \ge \mathcal{Q}^{sB,\gamma,\lambda}(\hat{\boldsymbol{\beta}})$ for all $\boldsymbol{\beta} \in B(\hat{\boldsymbol{\beta}}, \delta)$. From the convexity of $\mathcal{L}(\boldsymbol{\beta})$,

$$\mathcal{Q}^{sB,\gamma,\lambda}(\boldsymbol{\beta}) - \mathcal{Q}^{sB,\gamma,\lambda}(\hat{\boldsymbol{\beta}}) \ge \textstyle\sum_{j=1}^p v_j(\beta_j),$$

where

$$v_j(\beta_j) = -S_j(\hat{\boldsymbol{\beta}})(\beta_j - \hat{\beta}_j)/n + J^{\gamma,\lambda}(|\beta_j|) - J^{\gamma,\lambda}(|\hat{\beta}_j|).$$

First, if $|\hat{\beta}_j| > a(\lambda-\gamma)$, then

$$\begin{aligned} v_j(\beta_j) &= -\gamma\nu|\hat{\beta}_j|^{\nu-1}\text{sign}(\hat{\beta}_j)(\beta_j - \hat{\beta}_j) + \gamma|\beta_j|^\nu - \gamma|\hat{\beta}_j|^\nu \\ &\ge -\gamma\nu|\hat{\beta}_j|^{\nu-1}(|\beta_j| - |\hat{\beta}_j|) + \gamma|\beta_j|^\nu - \gamma|\hat{\beta}_j|^\nu \\ &\ge 0 \end{aligned}$$

for all $\beta_j \in B\big(\hat{\beta}_j, \delta_j\big)$, where $\delta_j = |\hat{\beta}_j| - a(\lambda-\gamma)$.

Next, if $\hat{\beta}_j = 0$, then

$$\begin{aligned} v_j(\beta_j) &\ge -|S_j(\hat{\boldsymbol{\beta}})\|\beta_j| + J^{\gamma,\lambda}(|\beta_j|) \\ &= |\beta_j|(-|S_j(\hat{\boldsymbol{\beta}})| - c^{\gamma,\lambda}|\beta_j|^\nu/(\nu+1) + \lambda) \ge 0 \end{aligned}$$

for all $\beta_j \in B\big(0, \delta_j\big)$, where

$$\delta_j = \big\{(\nu+1)(\lambda - |S_j(\hat{\boldsymbol{\beta}})|)/c^{\gamma,\lambda}\big\}^{1/\nu}.$$

Hence,

$$\mathcal{Q}^{sB,\gamma,\lambda}(\boldsymbol{\beta}) \ge \mathcal{Q}^{sB,\gamma,\lambda}(\hat{\boldsymbol{\beta}})$$

for all $\boldsymbol{\beta} \in B\big(\hat{\boldsymbol{\beta}}, \min_{1\le j\le p}\delta_j\big)$. This completes the proof. □

Lemma 2. *Assume the conditions in Theorem 3.3 hold then*

$$\mathbf{P}\big(\max_{q<j\le p}|S_j(\hat{\boldsymbol{\beta}}^{oB,\gamma})| < \lambda\big) \to 1,$$

and

$$\mathbf{P}\big(\min_{1\le j\le q}|\hat{\beta}_j^{oB,\gamma}| > a(\lambda-\gamma)\big) \to 1,$$

as $n \to \infty$.

Proof. Theorem 3.1 and (14) imply

$$\|\hat{\boldsymbol{\beta}}_1^{oB,\gamma} - \boldsymbol{\beta}_1^*\|_2 = o_p(\lambda). \tag{24}$$

From $(A1)$ and $(A2)$, we have

$$\begin{aligned} &\max_{q<j\le p}\big|S_j(\hat{\boldsymbol{\beta}}^{oB,\gamma})\big| \\ &\quad\le \max_{q<j\le p}\big|X_j'\mathbf{X}_1(\hat{\boldsymbol{\beta}}_1^{oB,\gamma} - \boldsymbol{\beta}_1^*)/n\big| + \max_{q<j\le p}|X_j{}'\boldsymbol{\varepsilon}/n| \\ &\quad\le b_0^{1/2}\tau_0^{1/2}\|\hat{\boldsymbol{\beta}}_1^{oB,\gamma} - \boldsymbol{\beta}_1^*\|_2 + \|\mathbf{X}_2'\boldsymbol{\varepsilon}/n\|_2. \end{aligned}$$

Since $E\|\mathbf{X}_2'\boldsymbol{\varepsilon}/n\|_2^2 \le (p-q)\tau_0\sigma_0^2/n = o(\lambda^2)$, from (24),

$$\max_{q<j\le p}|S_j(\hat{\boldsymbol{\beta}}^{oB,\gamma})| = o_p(\lambda).$$

On the other hand, from (24), $\min_{1\le j\le q}|\beta_j^*|/\lambda \to \infty$ implies

$$\begin{aligned} &\min_{1\le j\le q}|\hat{\beta}_j^{oB,\gamma}|/a(\lambda-\gamma) \\ &\quad\ge \min_{1\le j\le q}|\beta_j^*|/a(\lambda-\gamma) - \|\hat{\boldsymbol{\beta}}_1^{oB,\gamma} - \boldsymbol{\beta}_1^*\|_2/a(\lambda-\gamma) \\ &\quad\to \infty \end{aligned}$$

as $n \to \infty$. This completes the proof. □

Proof of Theorem 3.3. It is easy to see that (a) holds by Lemma 1 and 2. Hence, it suffices to prove (b). For this we will show

$$\mathbf{P}\big(\mathcal{Q}^{sB,\gamma,\lambda}(\hat{\boldsymbol{\beta}}^{oB,\gamma}) \le \inf_{\boldsymbol{\beta}}\mathcal{Q}^{sB,\gamma,\lambda}(\boldsymbol{\beta})\big) \to 1$$

as $n \to \infty$. For convenience, we omit the superscripts so that we write $\hat{\boldsymbol{\beta}}^{oB,\gamma} = \hat{\boldsymbol{\beta}}$ and $J^{\gamma,\lambda}(\cdot) = J(\cdot)$. From $(A1)$,

$$\mathcal{L}(\boldsymbol{\beta}) - \mathcal{L}(\hat{\boldsymbol{\beta}}) \ge \textstyle\sum_{j=1}^p\big\{-S_j(\hat{\boldsymbol{\beta}})(\beta_j - \hat{\beta}_j)/n + \rho_0(\beta_j - \hat{\beta}_j)^2/2\big\}.$$

Hence, from Lemma 2 and (19), it follows that

$$\mathcal{Q}^{sB,\gamma,\lambda}(\boldsymbol{\beta}) - \mathcal{Q}^{sB,\gamma,\lambda}(\hat{\boldsymbol{\beta}}) \ge \textstyle\sum_{j=1}^q w_j(\beta_j) + \sum_{j=q+1}^p v_j(\beta_j),$$

where

$$\begin{aligned} w_j(\beta_j) &= -\gamma\nu|\hat{\beta}_j|^{\nu-1}\text{sign}(\hat{\beta}_j)(\beta_j - \hat{\beta}_j) \\ &\quad + \rho_0(\beta_j - \hat{\beta}_j)^2/2 + J(|\beta_j|) - J(|\hat{\beta}_j|) \end{aligned}$$

and $v_j(\beta_j) = -o_p(\lambda)|\beta_j| + \rho_0|\beta_j|^2/2 + J(|\beta_j|)$.

First, consider the cases where $j \le q$. If $|\beta_j| \ge a(\lambda-\gamma)$,

$$w_j(\beta_j) \ge -\gamma\nu|\hat{\beta}_j|^{\nu-1}\text{sign}(\hat{\beta}_j)(\beta_j - \hat{\beta}_j) + \gamma(|\beta_j|^\nu - |\hat{\beta}_j|^\nu) \ge 0.$$

And if $|\beta_j| < a(\lambda-\gamma)$,

$$w_j(\beta_j) \geq -\gamma\nu|\hat{\beta}_j|^{\nu-1}(|\beta_j| - |\hat{\beta}_j|) + \rho_0(\beta_j - \hat{\beta}_j)^2/2 + J(|\beta_j|) - J(|\hat{\beta}_j|).$$

Since $J(a(\lambda-\gamma)) - J(|\hat{\beta}_j|) \geq \gamma\nu|\hat{\beta}_j|^{\nu-1}\{a(\lambda-\gamma) - |\hat{\beta}_j|\}$, it follows that

$$\begin{aligned} w_j(\beta_j) &\geq -\gamma\nu|\hat{\beta}_j|^{\nu-1}\{|\beta_j| - a(\lambda-\gamma)\} \\ &\quad + \rho_0(\beta_j - \hat{\beta}_j)^2/2 + J(|\beta_j|) - J(a(\lambda-\gamma)) \\ &\geq \rho_0(\beta_j - \hat{\beta}_j)^2/2 - J(a(\lambda-\gamma)). \end{aligned}$$

By triangular inequality, we have

$$|\beta_j - \hat{\beta}_j| \geq \min_{1\leq j\leq q}|\beta_j^*| - \|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 - a(\lambda-\gamma).$$

Hence, from (24), as $n \to \infty$,

$$|\beta_j - \hat{\beta}_j|/\{a(\lambda-\gamma)\} \to \infty$$

since $\min_{1\leq j\leq q}|\beta_j^*|/\lambda \to \infty$. On the other hand, $c^{\gamma,\lambda} = O(\lambda^{1-\nu})$ implies

$$\begin{aligned} &J(a(\lambda-\gamma))/\{a(\lambda-\gamma)\}^2 \\ &\quad = -c^{\gamma,\lambda}\{a(\lambda-\gamma)\}^{\nu-1}/(\nu+1) + \lambda/a(\lambda-\gamma) = O(1). \end{aligned}$$

Hence, $w_j(\beta_j) \geq 0$ for sufficiently large n. Next, consider the cases where $j > q$. If $|\beta_j| > a(\lambda-\gamma)$,

$$\begin{aligned} v_j(\beta_j) &> -o_p(\lambda)|\beta_j| + \rho_0|\beta_j|^2/2 \\ &> |\beta_j|\big(-o_p(\lambda) + \rho_0 a(\lambda-\gamma)/2\big) > 0 \end{aligned}$$

for all sufficiently large n. And if $|\beta_j| \leq a(\lambda-\gamma)$,

$$\begin{aligned} &v_j(\beta_j) \\ &\quad \geq -o_p(\lambda)|\beta_j| + J(|\beta_j|) \\ &\quad \geq |\beta_j|\big(-o_p(\lambda) - c^{\lambda,\gamma}\{a(\lambda-\gamma)\}^{\nu}/(\nu+1) + \lambda\big) \\ &\quad \geq |\beta_j|\big(-o_p(\lambda) - [\lambda - \gamma\nu\{a(\lambda-\gamma)\}^{\nu-1}]/(\nu+1) + \lambda\big) \\ &\quad \geq 0 \end{aligned}$$

for all sufficiently large n. Hence,

$$\mathcal{Q}^{sB,\gamma,\lambda}(\boldsymbol{\beta}) - \mathcal{Q}^{sB,\gamma,\lambda}(\hat{\boldsymbol{\beta}}) \geq 0$$

for all sufficiently large n which completes the proof. □

Received 9 December 2011

REFERENCES

[1] An, L. T. H. and Tao, P. D. (1997). Solving a class of linearly constrained indefinite quadratic problems by DC algorithms. *Journal of Global Optimization* **11** 253–285. MR1469128

[2] An, L. T. H. and Tao, P. D. (2005). The DC (Difference of Convex Functions) Programming and DCA revisited with DC models of real world nonconvex optimization problems. *Annals of Operations Research* **133** 23–46. MR2119311

[3] An, L. T. H., Le, M. H., Nguyen, V. V., and Tao, P. D. (2008). A DC programming approach for feature selection in support vector machines learning. *Advances in Data Analysis and Classification* **2** 259–278. MR2469770

[4] Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics* **24** 2350–2383. MR1425957

[5] Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455. MR1311089

[6] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics* **32** 407–499. MR2060166

[7] 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. MR1946581

[8] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics* **32** 928–961. MR2065194

[9] Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35** 109–148.

[10] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

[11] Huang, J., Horowitz, J. L., and Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *The Annals of Statistics* **36** 587–613. MR2396808

[12] Huang, J., Ma, S., and Zhang, C. H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica* **18** 1603–1618. MR2469326

[13] Kim, Y., Choi, H., and Oh, H. (2008). Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association* **103** 1656–1673. MR2510294

[14] Kim, Y. and Kwon, S. (2012). On the global optimality of nonconvex penalized estimators. *Biometrika* **99** 315–325.

[15] Knight, K. and Fu, W. (2000). Asymptotics for the LASSO-type estimators. *The Annals of Statistics* **28** 1356–1378. MR1805787

[16] Kwon, S. and Kim, Y. (2012). Large sample properties of the SCAD-penalized maximum likelihood estimation on high dimensions. *Statistica Sinica* **22** 629–953. MR2954355

[17] Park, M. and Hastie, T. (2007). l_1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society, Ser. B* **69** 659–667. MR2370074

[18] Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths. *The Annals of Statistics* **35** 1012–1030. MR2341696

[19] Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., Sheffield, V. C., and Stone, E. M. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proc. Nat. Acad. Sci.* **103** 14429–14434.

[20] Shen, X., Tseng, C., Zhang, X., and Wong, W. (2003). On psi-learning. *Journal of the American Statistical Association* **98** 724–734. MR2011686

[21] Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Ser. B* **58** 267–288. MR1379242

[22] Wang, H., Li, R., and Tsai, C-H. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94** 553–568. MR2410008

[23] Yuille, A. and Rangarajan, A. (2003). The concave-convex procedure. *Neural Computation* **15** 915–936.

[24] Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38** 894–942. MR2604701

[25] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Ser. B* **67** 301–320. MR2137327
[26] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101** 1418–1429. MR2279469
[27] Zou, H. and Zhang, H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics* **37** 1733–1751. MR2533470

Sunghoon Kwon
School of Statistics
University of Minnesota
Minneapolis, MN 55455
USA
E-mail address: shkwon0522@gmail.com

Yongdai Kim
Dept. of Statistics
Seoul National University
Seoul 151-742
Korea
E-mail address: ydkim0903@gmail.com

Hosik Choi
Dept. of Informational Statistics
and Institute of Basic Science
Hoseo University, Chungnam 336-795
Korea
E-mail address: choi.hosik@gmail.com