

*The Annals of Statistics*
2000, Vol. 28, No. 5, 1356–1378

# ASYMPTOTICS FOR LASSO-TYPE ESTIMATORS

BY KEITH KNIGHT[1] AND WENJIANG FU[2]

*University of Toronto and Michigan State University*

We consider the asymptotic behavior of regression estimators that minimize the residual sum of squares plus a penalty proportional to $\sum |\beta_j|^\gamma$ for some $\gamma > 0$. These estimators include the Lasso as a special case when $\gamma = 1$. Under appropriate conditions, we show that the limiting distributions can have positive probability mass at 0 when the true value of the parameter is 0. We also consider asymptotics for "nearly singular" designs.

**1. Introduction.** Consider the linear regression model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i = \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i \tag{1}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. random variables with mean 0 and variance $\sigma^2$. Without loss of generality, we will assume that the covariates are centered to have mean 0 and take $\hat{\beta}_0 = \overline{Y}$ in which case we can replace $Y_i$ in (1) by $Y_i - \overline{Y}$ and concentrate on estimating $\boldsymbol{\beta}$. Again without loss of generality, we will assume that $\overline{Y} = 0$.

We estimate $\boldsymbol{\beta}$ by minimizing the penalized least squares (LS) criterion

$$\sum_{i=1}^{n} \left(Y_i - \mathbf{x}_i^T \boldsymbol{\phi}\right)^2 + \lambda_n \sum_{j=1}^{p} |\phi_j|^\gamma \tag{2}$$

for a given $\lambda_n$ where $\gamma > 0$. Such estimators were called Bridge estimators by Frank and Friedman (1993) who introduced them as a generalization of ridge regression (which occurs for $\gamma = 2$). The special case when $\gamma = 1$ is related to the "Lasso" of Tibshirani (1996) (hence, the term "Lasso-type" in the title); in the case of wavelet regression, this approach to estimation is called basis pursuit [Chen, Donoho and Saunders (1999)]. Some other proposals for penalties are made in Fan and Li (1999). For $\gamma \leq 1$, the estimators minimizing (2) have the potentially attractive feature of being exactly 0 if $\lambda_n$ is sufficiently large, thus combining parameter estimation and model selection; indeed model selection methods that penalize by the number of nonzero parameters [such as AIC and BIC; Linhart and Zucchini (1986)] can be viewed as limiting cases of Bridge estimation as $\gamma \to 0$ since

$$\lim_{\gamma \downarrow 0} \sum_{j=1}^{p} |\phi_j|^\gamma = \sum_{j=1}^{p} I(\phi_j \neq 0).$$

Received June 1999; revised May 2000.

[1]Supported by the Natural Sciences and Engineering Research Council of Canada.

[2]Supported in part by grant R03OA83010 from the National Cancer Institute.

*AMS* 1991 *subject classifications*. Primary 62J05, 62J07; secondary 62E20, 60F05.

*Key words and phrases*. Penalized regression, Lasso, shrinkage estimation, epi-convergence in distribution.

For a given $\lambda_n$, we will denote the estimator minimizing (2) by $\hat{\boldsymbol{\beta}}_n$. Of course, $\lambda_n = 0$ corresponds to the ordinary LS estimator; this estimator will be denoted by $\hat{\boldsymbol{\beta}}_n^{(0)}$.

We will assume the following regularity conditions for the design,

$$C_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \to C, \tag{3}$$

where $C$ is a nonnegative definite matrix and

$$\frac{1}{n} \max_{1 \le i \le n} \mathbf{x}_i^T \mathbf{x}_i \to 0. \tag{4}$$

Typically in practice, the covariates are scaled so that the diagonal elements of $C_n$ (and hence those of $C$) are all identically 1.

The parametrization of the linear model (1) is unique if the matrix $C_n$ is nonsingular or, equivalently, the design matrix has full rank. It is worth noting, however, that a unique minimum to (2) may exist even if $C_n$ is singular; indeed, this is one of the benefits of this type of estimation. Define the "equivalence class"

$$\mathscr{B}_n = \{\zeta\colon \zeta = \boldsymbol{\beta} + \mathbf{v} \text{ where } C_n \mathbf{v} = \mathbf{0}\},$$

where $\boldsymbol{\beta}$ satisfies (1). When $C_n$ is singular, we could define a unique parametrization of (1) from the equivalence class $\mathscr{B}_n$ by defining $\boldsymbol{\beta}_0$ (for a given $\gamma > 0$) as

$$\boldsymbol{\beta}_0 = \operatorname{argmin}\left\{\sum_{j=1}^{p} |\zeta_j|^{\gamma}\colon \zeta \in \mathscr{B}_n\right\}.$$

However, this will not be pursued further here; we will assume that $C_n$ is nonsingular for all $n$.

An important class of designs is the class of "nearly singular" designs. For such designs, $C_n$ is nonsingular but may have one or more small eigenvalues (indicating the presence of collinearity among the covariates) such that (asymptotically) $C_n \to C$ where $C$ is singular. In practice, nearly singular designs can arise when many covariates are available, increasing the possibility of nearly linear dependencies between two or more covariates. These designs are considered in Section 5.

Under conditions (3) and (4) (with $C$ nonsingular), it is well-known that the LS estimator is consistent and that

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n^{(0)} - \boldsymbol{\beta}\right) \to_d N(\mathbf{0}, \sigma^2 C^{-1}).$$

In fact, conditions (3) and (4) can be weakened considerably without losing asymptotic normality of the LS estimator [Srivastava (1971)]; however, we will assume the existence of the limit $C$ in (3) throughout the paper. For the most part, we will assume that $C$ is nonsingular.

In Section 2, we discuss consistency and limiting distributions of Bridge estimators while in Section 3, we try to examine the small sample behavior by considering "local asymptotics." Asymptotics for bootstrapped Bridge estimation are considered in Section 4. In Section 5, we consider "nearly singular" designs as defined above. Finally, in Section 6 we try to tie up some other loose ends by considering Bridge estimation for singular designs as well as computation of Bridge estimators when $\gamma < 1$.

**2. Limiting distributions.** In this section, as well as in Section 3, we will assume that the matrix $C$ defined in (3) is nonsingular.

The limiting behavior of the Bridge estimator $\hat{\boldsymbol{\beta}}_n$ can be determined by studying the asymptotic behavior of the objective function (2). For example, to consider consistency of $\hat{\boldsymbol{\beta}}_n$, we will define the (random) function

$$Z_n(\boldsymbol{\phi}) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \mathbf{x}_i^T\boldsymbol{\phi}\right)^2 + \frac{\lambda_n}{n}\sum_{j=1}^{p}|\phi_j|^\gamma, \tag{5}$$

which is minimized at $\boldsymbol{\phi} = \hat{\boldsymbol{\beta}}_n$. The following result shows that $\hat{\boldsymbol{\beta}}_n$ is consistent provided $\lambda_n = o(n)$.

THEOREM 1. *If $C$ in (3) is nonsingular and $\lambda_n/n \to \lambda_0 \geq 0$, then $\hat{\boldsymbol{\beta}}_n \to_p$ argmin$(Z)$ where*

$$Z(\boldsymbol{\phi}) = (\boldsymbol{\phi} - \boldsymbol{\beta})^T C(\boldsymbol{\phi} - \boldsymbol{\beta}) + \lambda_0 \sum_{j=1}^{p}|\phi_j|^\gamma.$$

*Thus if $\lambda_n = o(n)$,* argmin$(Z) = \boldsymbol{\beta}$ *and so $\hat{\boldsymbol{\beta}}_n$ is consistent.*

PROOF. Define $Z_n$ as in (5). We need to show that

$$\sup_{\boldsymbol{\phi}\in K}|Z_n(\boldsymbol{\phi}) - Z(\boldsymbol{\phi}) - \sigma^2| \to_p 0 \tag{6}$$

for any compact set $K$ and that

$$\hat{\boldsymbol{\beta}}_n = O_p(1). \tag{7}$$

Under (6) and (7), we have

$$\text{argmin}(Z_n) \to_p \text{argmin}(Z).$$

For $\gamma \geq 1$, $Z_n$ is convex; thus (6) and (7) follow from the pointwise convergence in probability of $Z_n(\boldsymbol{\phi})$ to $Z(\boldsymbol{\phi}) + \sigma^2$ by applying standard results [Anderson and Gill (1982); Pollard (1991)]. For $\gamma < 1$, $Z_n$ is no longer convex but (6) follows easily. To prove (7), note that

$$Z_n(\boldsymbol{\phi}) \geq \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \mathbf{x}_i^T\boldsymbol{\phi}\right)^2 = Z_n^{(0)}(\boldsymbol{\phi})$$

for all $\boldsymbol{\phi}$. Since argmin$\left(Z_n^{(0)}\right) = O_p(1)$, it follows that argmin$(Z_n) = O_p(1)$. □

Even though $\lambda_n = o(n)$ is sufficient for consistency, we require that $\lambda_n$ grow more slowly for $\sqrt{n}$-consistency of the Bridge estimator. However, if $\lambda_n$ grows too slowly then $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ will have the same limiting distribution as $\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{(0)} - \boldsymbol{\beta})$. In fact, the rate of growth of $\lambda_n$ needed to get an "interesting" limiting distribution depends on whether $\gamma \geq 1$ or $\gamma < 1$. Theorem 2 below indicates that we need $\lambda_n = O(\sqrt{n})$ for $\sqrt{n}$-consistency for $\gamma \geq 1$; Theorem 3 suggests $\lambda_n = O(n^{\gamma/2})$ is necessary for $\gamma < 1$. [In fact, $\lambda_n = O(\sqrt{n})$ suffices for $\gamma < 1$.]

THEOREM 2. *Suppose that* $\gamma \geq 1$. *If* $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$ *and* $C$ *is nonsingular then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to_d \operatorname{argmin}(V),$$

*where*

$$V(\mathbf{u}) = -2\mathbf{u}^T\mathbf{W} + \mathbf{u}^T C\mathbf{u} + \lambda_0 \sum_{j=1}^{p} u_j \operatorname{sgn}(\beta_j)|\beta_j|^{\gamma-1}$$

*if* $\gamma > 1$,

$$V(\mathbf{u}) = -2\mathbf{u}^T\mathbf{W} + \mathbf{u}^T C\mathbf{u} + \lambda_0 \sum_{j=1}^{p} [u_j \operatorname{sgn}(\beta_j) I(\beta_j \neq 0) + |u_j| I(\beta_j = 0)]$$

*if* $\gamma = 1$, *and* $\mathbf{W}$ *has a* $N(\mathbf{0}, \sigma^2 C)$ *distribution.*

PROOF. Define $V_n(\mathbf{u})$ by

$$V_n(\mathbf{u}) = \sum_{i=1}^{n} \left[ \left(\varepsilon_i - \mathbf{u}^T\mathbf{x}_i/\sqrt{n}\right)^2 - \varepsilon_i^2 \right] + \lambda_n \sum_{j=1}^{p} \left[ |\beta_j + u_j/\sqrt{n}|^\gamma - |\beta_j|^\gamma \right]$$

[where $\mathbf{u} = (u_1, \ldots, u_p)^T$] and note that $V_n$ is minimized at $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$. First note that

$$\sum_{i=1}^{n} \left[ \left(\varepsilon_i - \mathbf{u}^T\mathbf{x}_i/\sqrt{n}\right)^2 - \varepsilon_i^2 \right] \to_d -2\mathbf{u}^T\mathbf{W} + \mathbf{u}^T C\mathbf{u}$$

with finite-dimensional convergence holding trivially. If $\gamma > 1$ then

$$\lambda_n \sum_{j=1}^{p} \left[ |\beta_j + u_j/\sqrt{n}|^\gamma - |\beta_j|^\gamma \right] \to \lambda_0 \sum_{j=1}^{p} u_j \operatorname{sgn}(\beta_j)|\beta_j|^{\gamma-1},$$

while for $\gamma = 1$, we have

$$\lambda_n \sum_{j=1}^{p} [|\beta_j + u_j/\sqrt{n}| - |\beta_j|] \to \lambda_0 \sum_{j=1}^{p} [u_j \operatorname{sgn}(\beta_j) I(\beta_j \neq 0) + |u_j| I(\beta_j = 0)].$$

Thus $V_n(\mathbf{u}) \to_d V(\mathbf{u})$ (as defined above) with the finite-dimensional convergence holding trivially. Since $V_n$ is convex and $V$ has a unique minimum, it follows [Geyer (1996)] that

$$\operatorname{argmin}(V_n) = \sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to_d \operatorname{argmin}(V).$$

Note that when $\lambda_0 = 0$, $\operatorname{argmin}(V) = C^{-1}\mathbf{W} \sim N(\mathbf{0}, \sigma^2 C^{-1})$. □

Proponents of ridge regression (that is, $\gamma = 2$) may be disappointed with the conclusion of Theorem 2 as it gives

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to_d C^{-1}(\mathbf{W} - \lambda_0\boldsymbol{\beta}) \sim N(-\lambda_0 C^{-1}\boldsymbol{\beta}, \sigma^2 C^{-1}),$$

which suggests that ridge estimation is inferior to ordinary LS estimation. However, the asymptotic perspective used here is somewhat unfair to ridge estimation; see Theorem 4 in Section 3 for a more flattering asymptotic perspective of ridge estimation. However, Theorem 2 does illustrate that for $\gamma > 1$, the amount of shrinkage towards 0 increases with the magnitude of the parameter being estimated; thus, for "large" parameters, the bias of their estimators for $\gamma > 1$ may be unacceptably large.

THEOREM 3. *Suppose that* $\gamma < 1$. *If* $\lambda_n / n^{\gamma/2} \to \lambda_0 \geq 0$ *then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to_d \operatorname{argmin}(V),$$

*where*

$$V(\mathbf{u}) = -2\mathbf{u}^T\mathbf{W} + \mathbf{u}^T C\mathbf{u} + \lambda_0 \sum_{j=1}^{p} |u_j|^\gamma I(\beta_j = 0).$$

PROOF. The proof is similar to that of Theorem 2; however, there are some added complexities due to the nonconvexity of the objective function. Define

$$V_n(\mathbf{u}) = \sum_{i=1}^{n} \left[ \left(\varepsilon_i - \mathbf{u}^T\mathbf{x}_i/\sqrt{n}\right)^2 - \varepsilon_i^2 \right] + \lambda_n \sum_{j=1}^{p} \left[ |\beta_j + u_j/\sqrt{n}|^\gamma - |\beta_j|^\gamma \right]. \tag{8}$$

Since $\lambda_n = O(n^{\gamma/2}) = o(\sqrt{n})$, it follows that

$$\lambda_n[|\beta_j + u_j/\sqrt{n}|^\gamma - |\beta_j|^\gamma] \to 0$$

if $\beta_j \neq 0$. Thus

$$\lambda_n \sum_{j=1}^{p} \left[ |\beta_j + u_j/\sqrt{n}|^\gamma - |\beta_j|^\gamma \right] \to \lambda_0 \sum_{j=1}^{p} |u_j|^\gamma I(\beta_j = 0)$$

and the convergence is uniform over $\mathbf{u}$ in compact sets. It follows then that

$$V_n(\cdot) \to_d V(\cdot)$$

on the space of functions topologized by uniform convergence on compact sets. To prove that $\operatorname{argmin}(V_n) \to_d \operatorname{argmin}(V)$, it suffices to show that $\operatorname{argmin}(V_n) = O_p(1)$ [Kim and Pollard (1990)]. However, note that

$$\begin{aligned} V_n(\mathbf{u}) &\geq \sum_{i=1}^{n}\left[(\varepsilon_i - \mathbf{u}^T\mathbf{x}_i/\sqrt{n})^2 - \varepsilon_i^2\right] - \lambda_n \sum_{j=1}^{p} |u_j/\sqrt{n}|^\gamma \\ &\geq \sum_{i=1}^{n}\left[(\varepsilon_i - \mathbf{u}^T\mathbf{x}_i/\sqrt{n})^2 - \varepsilon_i^2\right] - (\lambda_0 + \delta) \sum_{j=1}^{p} |u_j|^\gamma \\ &= V_n^{(l)}(\mathbf{u}) \end{aligned}$$

for all $\mathbf{u}$ and $n$ sufficiently large. Since the quadratic terms in $V_n^{(l)}$ grow faster than the $|u_j|^\gamma$ terms, it follows that $\operatorname{argmin}(V_n^{(l)}) = O_p(1)$; hence, it follows that $\operatorname{argmin}(V_n) = O_p(1)$. Since $\operatorname{argmin}(V)$ is unique with probability 1, the conclusion follows. □

The conclusion of Theorem 3 is quite interesting. It suggests that if $\gamma < 1$, we can estimate nonzero regression parameters at the usual rate without asymptotic bias while shrinking the estimates of zero regression parameters to 0 with positive probability. This is in contrast to what happens when $\gamma \geq 1$; Theorem 2 indicates that nonzero parameters are estimated with some asymptotic bias if $\lambda_0 > 0$.

As an alternative to the conditions on $\lambda_n$ given in Theorem 3, we can also consider what happens if $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$ while $\lambda_n/n^{\gamma/2} \to \infty$. Suppose that $\beta_1, \ldots, \beta_r$ are nonzero while $\beta_{r+1}, \ldots, \beta_p$ are zero. Then defining $V_n(\mathbf{u})$ as in (8), it follows that $V_n(\mathbf{u}) \to_d V(\mathbf{u})$ where

$$V(\mathbf{u}) = \begin{cases} -2\mathbf{u}^T\mathbf{W} + \mathbf{u}^T C\mathbf{u} + \gamma\lambda_0 \sum_{j=1}^{r}[u_j|\beta_j|^\gamma/\beta_j], & \text{if } u_{r+1} = \cdots = u_p = 0, \\ \infty, & \text{otherwise.} \end{cases}$$

[In fact, since $V$ can be infinite, we can no longer define convergence of $V_n$ to $V$ via uniform convergence on compact sets but instead define it via epiconvergence which allows for extended real-valued functions; see Geyer (1994, 1996), Pflug (1995) for more details on epiconvergence.] Applying the arguments given in the proof of Theorem 3, it follows that $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to_d \operatorname{argmin}(V)$ where the last $(p - r)$ elements of $\operatorname{argmin}(V)$ are exactly 0. This result suggests that we could achieve the best of both worlds (at least asymptotically) by taking $\lambda_n \sim \lambda_0 n^{\alpha/2}$ where $\gamma < \alpha < 1$. However, this latter formulation does not really capture what happens for finite samples.

When some of the $\beta_j$'s are exactly 0, the limiting distributions (as given by Theorems 2 and 3) put positive probability at 0 when $\gamma \leq 1$. We will illustrate this in the case when $\gamma = 1$. Suppose that $\beta_1, \ldots, \beta_r$ are nonzero and $\beta_{r+1} = \cdots = \beta_p = 0$. In this case,

$$V(\mathbf{u}) = -2\mathbf{u}^T\mathbf{W} + \mathbf{u}^T C\mathbf{u} + \lambda_0 \sum_{j=1}^{r} u_j \operatorname{sgn}(\beta_j) + \lambda_0 \sum_{j=r+1}^{p} |u_j|$$

Now rewrite the matrix $C$, $\mathbf{W}$ and $\mathbf{u}$ as follows:

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix},$$

where $C_{11}$ is $r \times r$, $C_{22}$ is $(p-r) \times (p-r)$ and $C_{21} = C_{12}^T$;

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{pmatrix},$$

$$\mathbf{u} = \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{pmatrix},$$

where $\mathbf{W}_1$ and $\mathbf{u}_1$ are $r$-vectors. If $V(\mathbf{u})$ is minimized at $\mathbf{u}_2 = \mathbf{0}$ then it follows that

$$C_{11}\mathbf{u}_1 - \mathbf{W}_1 = -\frac{\lambda_0}{2}\begin{pmatrix} \operatorname{sgn}(\beta_1) \\ \vdots \\ \operatorname{sgn}(\beta_r) \end{pmatrix} = -\frac{\lambda_0}{2}s(\beta) \tag{9}$$

and

$$-\frac{\lambda_0}{2}\mathbf{1} \le C_{21}\mathbf{u}_1 - \mathbf{W}_2 \le \frac{\lambda_0}{2}\mathbf{1}, \tag{10}$$

where $\mathbf{1}$ is a vector of 1's and the inequalities are interpreted coordinatewise. Solving for $\mathbf{u}_1$ in (9), we get

$$\mathbf{u}_1 = C_{11}^{-1}(\mathbf{W}_1 - \lambda_0 s(\boldsymbol{\beta})/2)$$

and substituting into (10), it follows that $\mathbf{u}_2 = \mathbf{0}$ if

$$-\frac{\lambda_0}{2}\mathbf{1} \le C_{21}C_{11}^{-1}(\mathbf{W}_1 - \lambda_0 s(\boldsymbol{\beta})/2) - \mathbf{W}_2 \le \frac{\lambda_0}{2}\mathbf{1}$$

In the case where $\beta_1 = \beta_2 = \cdots = \beta_p = 0$, this reduces to

$$-\frac{\lambda_0}{2}\mathbf{1} \le \mathbf{W} \le \frac{\lambda_0}{2}\mathbf{1}.$$

Note that this same rationale applies for finite samples as well; for example, $\hat{\boldsymbol{\beta}}_n = \mathbf{0}$ if and only if $-\lambda_n\mathbf{1} \le 2\sum_i Y_i\mathbf{x}_i \le \lambda_n\mathbf{1}$.

EXAMPLE 1. Consider a quadratic regression model

$$\begin{aligned} Y_i &= \beta_0 + \beta_1\frac{(x_i - a_n^{(1)})}{s_n^{(1)}} + \beta_2\frac{(x_i^2 - a_n^{(2)})}{s_n^{(2)}} + \varepsilon_i \\ &= \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \varepsilon_i \quad \text{for } i = 1, \ldots, n, \end{aligned}$$

where the $x_i$'s are uniformly distributed over the interval $[0, 1]$ with

$$a_n^{(1)} = \frac{1}{n}\sum_{i=1}^{n} x_i \to \frac{1}{2},$$

$$a_n^{(2)} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \to \frac{1}{3}$$

$$s_n^{(1)} = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - a_n^{(1)})^2\right)^{1/2} \to \frac{1}{\sqrt{12}},$$

and

$$s_n^{(2)} = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i^2 - a_n^{(2)})^2\right)^{1/2} \to \frac{2}{3\sqrt{5}}.$$

In this case,

$$C_n \to C = \begin{pmatrix} 1 & \sqrt{15/16} \\ \sqrt{15/16} & 1 \end{pmatrix}.$$

We define estimators $\hat{\beta}_{nl}$, $\hat{\beta}_{n2}$ to minimize

$$\sum_{i=1}^{n}(Y_i - \overline{Y} - \phi_1 z_{1i} - \phi_2 z_{2i})^2 + \lambda_n[|\phi_1|^\gamma + |\phi_2|^\gamma].$$

In this example, we will consider the cases $\gamma = 1$ and $\gamma = 1/2$ with $\lambda_n/n^{\gamma/2} \to \lambda_0 > 0$ and $\beta_1 > 0$, $\beta_2 = 0$. Then

$$\sqrt{n}\begin{pmatrix} \hat{\beta}_{nl} - \beta_1 \\ \hat{\beta}_{n2} \end{pmatrix} \to_d \operatorname{argmin}(V),$$

where

$$V(u_1, u_2) = -2(u_1 W_1 + u_2 W_2) + u_1^2 + \frac{\sqrt{15}}{2}u_1 u_2 + u_2^2 + \lambda_0[u_1 + |u_2|]$$

for $\gamma = 1$,

$$V(u_1, u_2) = -2(u_1 W_1 + u_2 W_2) + u_1^2 + \frac{\sqrt{15}}{2}u_1 u_2 + u_2^2 + \lambda_0|u_2|^{1/2}$$

for $\gamma = 1/2$, and $(W_1, W_2)$ is a zero mean bivariate Normal random vector with covariance matrix $\sigma^2 C$.

For $\gamma = 1$ (assuming that $\beta_1 > 0$ and $\beta_2 = 0$), it is fairly straightforward to evaluate argmin$(V)$ explicitly; letting $(\widehat{U}_1, \widehat{U}_2) = \operatorname{argmin}(V)$, we have three

cases to consider, depending on the value of

$$\tau(W_1, W_2, \lambda_0) = \frac{\sqrt{15}}{2} W_1 - 2W_2 - \frac{\sqrt{15}}{4}\lambda_0.$$

1. If $|\tau(W_1, W_2, \lambda_0)| \leq \lambda_0$ then

$$\widehat{U}_1 = W_1 - \frac{\lambda_0}{2},$$

$$\widehat{U}_2 = 0.$$

2. If $\tau(W_1, W_2, \lambda_0) < -\lambda_0$ then

$$\widehat{U}_1 = 16W_1 - 4\sqrt{15}W_2 + (2\sqrt{15} - 8)\lambda_0,$$

$$\widehat{U}_2 = 16W_2 - 4\sqrt{15}W_1 + (2\sqrt{15} - 8)\lambda_0.$$

3. If $\tau(W_1, W_2, \lambda_0) > \lambda_0$ then

$$\widehat{U}_1 = 16W_1 - 4\sqrt{15}W_2 - (2\sqrt{15} + 8)\lambda_0,$$

$$\widehat{U}_2 = 16W_2 - 4\sqrt{15}W_1 + (2\sqrt{15} + 8)\lambda_0.$$

For $\gamma = 1/2$, an exact representation of the distribution of $(\widehat{U}_1, \widehat{U}_2)$ is more difficult to obtain but this distribution can be easily simulated.

Tables 1 and 2 give the means, variances and correlations of $\widehat{U}_1$ and $\widehat{U}_2$ as well as $P(\widehat{U}_2 = 0)$ for various values of $\lambda_0$ (scaled by $\sigma$). As one might expect, the asymptotic variances of $\hat{\beta}_{n1}$ and $\hat{\beta}_{n2}$ decrease as $\lambda_0$ increases. On the other hand, the asymptotic bias of $\hat{\beta}_{n1}$ becomes increasingly negative as $\lambda_0$ increases while the asymptotic bias of $\hat{\beta}_{n2}$ increases away from zero then decreases to 0 as $\lambda_0$ increases.

Figures 1, 2, and 3 show scatterplots of random samples of 500 drawn from the limiting distributions for the LS estimator ($\lambda_0 = 0$), $\lambda_0 = 1.0$ for $\gamma = 1$, and $\lambda_0 = 0.5$ for $\gamma = 1/2$, respectively (with $\sigma^2 = 1$). [These values of $\lambda_0$ for the different values of $\gamma$ give approximately equal values of $P(\widehat{U}_2 = 0)$.] To facilitate comparison, the same values of $(W_1, W_2)$ were used to generate the three

TABLE 1

*Properties of the distribution of* argmin$(V)$ *for* $\gamma = 1$ *and various values of* $\lambda_0$

| $\frac{\lambda_0}{\sigma}$ | $\frac{E(\widehat{U}_1)}{\sigma}$ | $\frac{E(\widehat{U}_2)}{\sigma}$ | $\frac{\mathrm{Var}(\widehat{U}_1)}{\sigma^2}$ | $\frac{\mathrm{Var}(\widehat{U}_2)}{\sigma^2}$ | $\mathrm{Corr}(\widehat{U}_1, \widehat{U}_2)$ | $P(\widehat{U}_2 = 0)$ |
|---|---|---|---|---|---|---|
| 0.0 | 0.00 | 0.00 | 16.00 | 16.00 | −0.968 | 0.000 |
| 0.1 | −0.68 | 0.65 | 11.90 | 11.62 | −0.957 | 0.156 |
| 0.2 | −1.14 | 1.07 | 8.89 | 8.49 | −0.944 | 0.290 |
| 0.5 | −1.71 | 1.50 | 6.16 | 5.53 | −0.915 | 0.488 |
| 1.0 | −1.93 | 1.47 | 5.78 | 5.10 | −0.909 | 0.525 |
| 2.0 | −2.33 | 1.37 | 5.36 | 4.71 | −0.901 | 0.550 |
| 5.0 | −3.51 | 1.04 | 4.40 | 3.63 | −0.876 | 0.624 |

TABLE 2
*Properties of the distribution of* $\operatorname{argmin}(V)$ *for* $\gamma = 0.5$ *and various values of* $\lambda_0$

| $\frac{\lambda_0}{\sigma^{3/2}}$ | $\frac{E(\widehat{U}_1)}{\sigma}$ | $\frac{E(\widehat{U}_2)}{\sigma}$ | $\frac{\operatorname{Var}(\widehat{U}_1)}{\sigma^2}$ | $\frac{\operatorname{Var}(\widehat{U}_2)}{\sigma^2}$ | $\operatorname{Corr}(\widehat{U}_1, \widehat{U}_2)$ | $P(\widehat{U}_2 = 0)$ |
|---|---|---|---|---|---|---|
| 0.0 | 0.00 | 0.00 | 16.00 | 16.00 | −0.968 | 0.000 |
| 0.1 | 0.00 | 0.00 | 14.86 | 14.78 | −0.966 | 0.193 |
| 0.2 | 0.00 | 0.00 | 13.73 | 13.57 | −0.963 | 0.303 |
| 0.5 | 0.00 | 0.00 | 10.77 | 10.41 | −0.952 | 0.529 |
| 1.0 | 0.00 | 0.00 | 7.06 | 6.46 | −0.926 | 0.745 |
| 2.0 | 0.00 | 0.00 | 3.09 | 2.21 | −0.821 | 0.930 |
| 5.0 | 0.00 | 0.00 | 1.05 | 0.04 | −0.197 | 0.999 |

scatterplots. These scatterplots illustrate the effect of Bridge estimation relative to LS estimation from an asymptotic point of view. In LS estimation, the strong correlation between the two variables means that overestimation of $\beta_1$ is generally accompanied by underestimation of $\beta_2$ (and vice versa); moreover, this effect holds regardless of the true values of $\beta_1$ and $\beta_2$. In contrast, Lasso estimation ($\gamma = 1$) compensates for underestimation of $\beta_1$ by overestimation of $\beta_2$ but effectively sets the estimate of $\beta_2$ to zero if $\beta_1$ is overestimated. For $\gamma = 1/2$, the shrinkage to zero in the estimation of $\beta_2$ is more selective and "larger" estimates are essentially unchanged from their corresponding LS estimates. Also note that there is a "no man's land" in the distribution of $\widehat{U}_2$ when $\gamma = 1/2$; for each $\lambda_0$, there is an open interval $I(\lambda_0) = (0, c(\lambda_0))$ [with $c(\lambda_0) > 0$] such that $P[|\widehat{U}_2|] \notin I(\lambda_0)] = 1$. For $\lambda_0 = 1/2$, $c(\lambda_0) \approx 0.86$. □

How well do the asymptotic distributions approximate finite sample distributions? There are a number of factors involved including the accuracy of

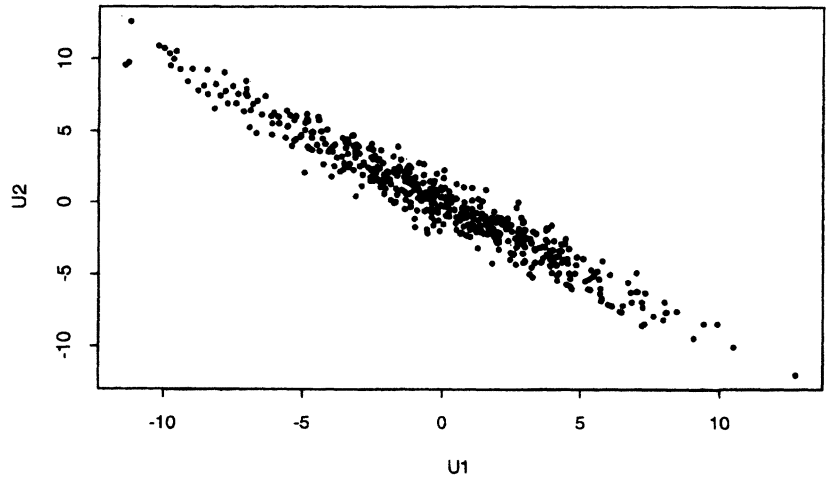


FIG. 1. *Sample of* 500 *from the limiting distribution of the LS estimator in Example* 1.

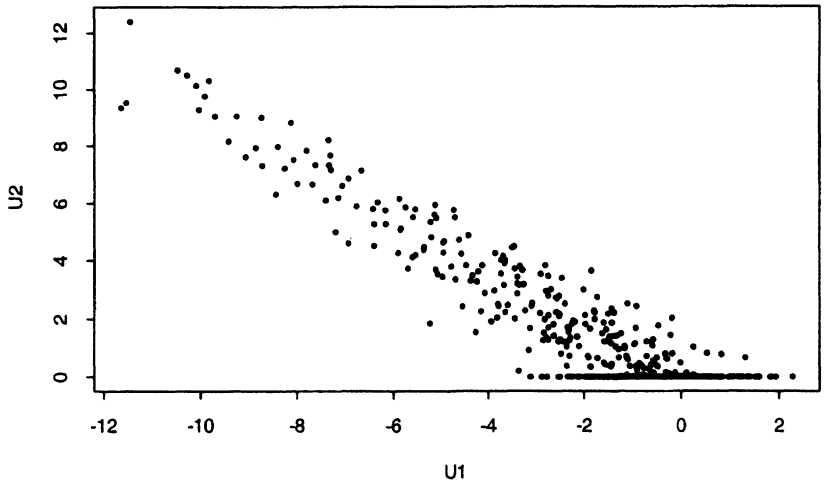


FIG. 2. *Sample of* 500 *from the limiting distribution of the Bridge estimator in Example* 1 *with* $\gamma = 1$ *and* $\lambda_0 = 1$. *The probability that* $\widehat{U}_2$ *is strictly less than* 0 *is approximately* $4.1 \times 10^{-5}$, *which explains the absence of negative* $\widehat{U}_2$ *values.*

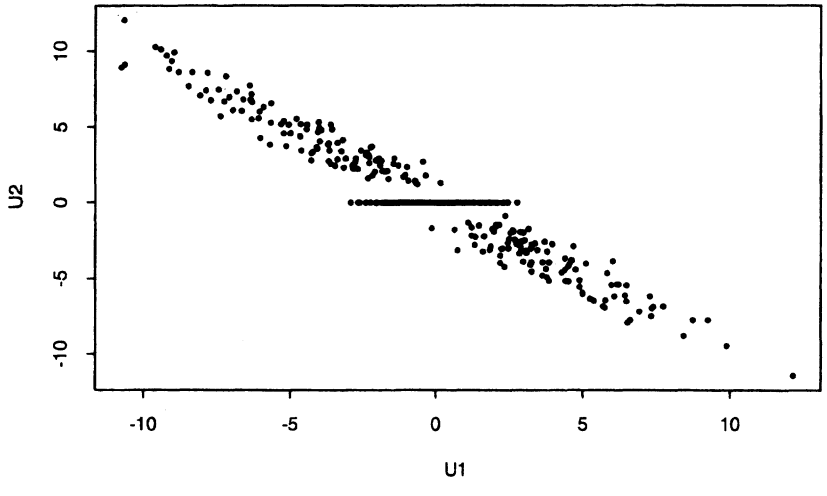


FIG. 3. *Sample of* 500 *from the limiting distribution of the Bridge estimator in Example* 1 *with* $\gamma = 1/2$ *and* $\lambda_0 = 1/2$.

normal approximations; however, the key factor here would seem to be the extent to which the asymptotic penalty in $V(\mathbf{u})$ (defined in Theorems 2 and 3) approximates the true penalty term. For example, when $\gamma \leq 1$, these approximations may not be particularly good for finite samples as the function $|x|^\gamma$ is not particularly smooth when $x$ is close to 0. This is addressed to some extent in the next section.

**3. Local asymptotics and small parameters.** A distinguishing feature of Bridge estimation for $\gamma \leq 1$ is the possibility of obtaining exact 0 parameter estimates. In the previous section, we showed that the limiting distributions have positive mass at 0 when the true parameter value is 0 but are absolutely continuous (with respect to Lebesgue measure) otherwise. In this section, we will try to illustrate how this "exact 0" phenomenon can occur in finite samples when the true parameter is small but nonzero.

To do this, we will assume that we have a triangular array of observations. That is, define

$$Y_{ni} = \boldsymbol{\beta}_n^T \mathbf{x}_{ni} + \varepsilon_{ni} \quad \text{for } i = 1, \ldots, n, \tag{11}$$

where for each $n$, $\varepsilon_{n1}, \ldots, \varepsilon_{nn}$ are i.i.d. random variables with mean 0 and variance $\sigma^2$. We assume that the $\mathbf{x}_{ni}$'s satisfy the conditions

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_{ni} \mathbf{x}_{ni}^T \to C \tag{12}$$

for some positive definite matrix $C$ and

$$\frac{1}{n} \max_{1 \leq i \leq n} \mathbf{x}_{ni}^T \mathbf{x}_{ni} \to 0; \tag{13}$$

these are the obvious analogues of (3) and (4).

Suppose that $\boldsymbol{\beta}_n = \boldsymbol{\beta} + t/\sqrt{n}$ and define $\hat{\boldsymbol{\beta}}_n$ to minimize

$$\sum_{i=1}^{n} (Y_{ni} - \boldsymbol{\phi}^T \mathbf{x}_{ni})^2 + \lambda_n \sum_{j=1}^{p} |\phi_j|. \tag{14}$$

This formulation allows us to examine the asymptotic properties of Bridge estimators when one or more of the regression parameters are close to 0 but nonzero. The idea here is to get a hint of the small sample behavior of Bridge estimation.

THEOREM 4. *Assume the model* (11) *with* $\boldsymbol{\beta}_n = \boldsymbol{\beta} + t/\sqrt{n}$ *and assume that* (12) *and* (13) *are satisfied. Let* $\hat{\boldsymbol{\beta}}_n$ *minimize* (14) *for some* $\gamma > 1$.

(a) *If* $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$ *then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n) \to_d \operatorname{argmin}(V),$$

*where*

$$V(\mathbf{u}) = -2\mathbf{u}^T \mathbf{W} + \mathbf{u}^T C \mathbf{u} + \lambda_0 \sum_{j=1}^{p} u_j \operatorname{sgn}(\beta_j) |\beta_j|^{\gamma - 1}.$$

(b) *If* $\boldsymbol{\beta} = \mathbf{0}$ *and* $\lambda_n/n^{\gamma/2} \to \lambda_0 \geq 0$ *then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n) \to_d \operatorname{argmin}(V),$$

*where*

$$V(\mathbf{u}) = -2\mathbf{u}^T\mathbf{W} + \mathbf{u}^T C\mathbf{u} + \lambda_0 \sum_{j=1}^{p} |u_j + t_j|^\gamma.$$

THEOREM 5. *Assume the model* (11) *with* $\boldsymbol{\beta}_n = \boldsymbol{\beta} + \mathbf{t}/\sqrt{n}$ *and assume that* (12) *and* (13) *are satisfied. Suppose that* $\hat{\boldsymbol{\beta}}_n$ *minimizes* (14) *for* $\gamma \leq 1$ *where* $\lambda_n/n^{\gamma/2} \to \lambda_0 \geq 0$.

(a) *For* $\gamma = 1$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n) \to_d \operatorname{argmin}(V)$$

*where*

$$V(\mathbf{u}) = -2\mathbf{u}^T\mathbf{W} + \mathbf{u}^T C\mathbf{u} + \lambda_0 \sum_{j=1}^{p} [u_j \operatorname{sgn}(\beta_j) I(\beta_j \neq 0) + |u_j + t_j| I(\beta_j = 0)]$$

(b) *For* $\gamma < 1$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n) \to_d \operatorname{argmin}(V),$$

*where*

$$V(\mathbf{u}) = -2\mathbf{u}^T\mathbf{W} + \mathbf{u}^T C\mathbf{u} + \lambda_0 \sum_{j=1}^{p} |u_j + t_j|^\gamma I(\beta_j = 0).$$

The proofs of Theorems 4 and 5 are essentially the same as those of Theorems 2 and 3. Theorem 4 suggests that the advantages of using a penalty with $\gamma > 1$ are limited to situations where all the regression parameters are relatively small (compared to $n$); for example, for ridge estimation ($\gamma = 2$), part (b) of Theorem 4 gives

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \mathbf{t}/\sqrt{n}) &\to_d (C + \lambda_0 I)^{-1}(\mathbf{W} - \lambda_0 \mathbf{t}) \\ &\sim N(-\lambda_0((C + \lambda_0 I)^{-1}\mathbf{t}, \sigma^2(C + \lambda_0 I)^{-1} C(C + \lambda_0 I)^{-1}),\end{aligned}$$

which suggests that by choosing $\lambda_0$ judiciously, we could make the mean square error of $\mathbf{x}^T\hat{\boldsymbol{\beta}}_n$ smaller than that of $\mathbf{x}^T\hat{\boldsymbol{\beta}}_n^{(0)}$. On the other hand, if one or more of the parameters is "large" then part (a) of Theorem 4 indicates that the bias suggested by Theorem 2 would still persist.

In contrast, when $\gamma \leq 1$, Theorem 5 suggests that "small" parameters may be estimated as exactly 0 even when "large" parameters are present. For example, suppose that $\beta_{nj} = t_j/\sqrt{n}$ and take $\gamma \leq 1$; then the limiting distribution of $\sqrt{n}(\hat{\beta}_{nj} - t_j/\sqrt{n})$ (given by Theorem 5) puts positive probability mass at $-t_j$ and so the limiting distribution of $\sqrt{n}\hat{\beta}_{nj}$ puts positive probability mass at 0, irrespective of the values of the other parameters. Thus, Theorem 5 indicates that "small" parameters may be estimated as exactly 0 in finite samples even when "large" parameters are present.

**4. Bootstrapping.** Attaching standard error estimates to Bridge parameter estimates is nontrivial especially when $\gamma \leq 1$. For the Lasso ($\gamma = 1$), Tibshirani (1996) gives an approximation of the covariance matrix of the estimators. However, his approximation leads to standard error estimates of 0 when the estimate is 0, which is clearly unsatisfactory; Osborne, Presnell and Turlach (1998) give an alternative approximation that leads to apparently more satisfactory standard error estimates. However, these approximations to the covariance matrix implicitly assume that the estimators are approximately linear transformations, which is clearly not the case when $\gamma \leq 1$. An alternative approach to obtaining standard error estimates is to use the bootstrap.

In regression models, there are effectively two approaches to bootstrapping depending on whether the design is considered fixed or random.

1. (Random design). We draw a bootstrap sample $(Y_1^*, \mathbf{x}_1^*), \ldots, (Y_n^*, \mathbf{x}_n^*)$ with replacement from $\{(Y_1, \mathbf{x}_i), \ldots, (Y_n, \mathbf{x}_n)\}$.
2. (Fixed design). The bootstrap sample is $(Y_1^*, \mathbf{x}_1), \ldots, (Y_n^*, \mathbf{x}_n)$ where

$$Y_i^* = \overline{Y} + \tilde{\boldsymbol{\beta}}_n^T \mathbf{x}_i + \varepsilon_i^* \quad \text{for } i = 1, \ldots, n$$

with $\varepsilon_1^*, \ldots, \varepsilon_n^*$ sampled with replacement from "residuals" $\{e_1, \ldots, e_n\}$ and $\tilde{\boldsymbol{\beta}}_n$ some estimator of $\boldsymbol{\beta}$ (not necessarily a Bridge estimator).

Using the bootstrap sample, we can then obtain a bootstrap Bridge estimator of $\boldsymbol{\beta}$ (call it $\hat{\boldsymbol{\beta}}_n^*$) by minimizing an appropriate version of (2) for some $\gamma$ and $\lambda_n$. The idea is that the bootstrap distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}}_n^* - \tilde{\boldsymbol{\beta}}_n)$ should approximate the sampling distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n)$ where $\hat{\boldsymbol{\beta}}_n$ is the Bridge estimator minimizing (2).

The asymptotics of approach (2) are quite simple. Assume that $\sqrt{n}(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to_d \mathbf{U}$ and the set of residuals sampled from has mean 0. For each bootstrap sample, we will centre the $Y_i^*$'s by their sample mean. Define.

$$V_n^*(\mathbf{u}) = \sum_{i=1}^{n} \left[ (\varepsilon_i^* - \mathbf{u}^T \mathbf{x}_i / \sqrt{n})^2 - (\varepsilon_i^*)^2 \right] + \lambda_n \sum_{j=1}^{p} \left[ |\tilde{\beta}_{nj} + u_j / \sqrt{n}|^\gamma - |\tilde{\beta}_{nj}|^\gamma \right].$$

Conditional on $\tilde{\boldsymbol{\beta}}_n$, the randomness of $V_n^*$ comes from the bootstrap sampling producing the $\varepsilon_i^*$'s. The idea here is exactly the same as before: if $V_n^*$ converges to some $V^*$ then the bootstrap distribution of $\operatorname{argmin}(V_n^*) = \sqrt{n}(\hat{\boldsymbol{\beta}}_n^* - \tilde{\boldsymbol{\beta}}_n)$ should converge (in some sense) to that of $\operatorname{argmin}(V^*)$. What complicates matters is the fact that there are two layers of randomness: one due to the original sample (reflected through $\tilde{\boldsymbol{\beta}}_n$) and one due to the bootstrap sampling.

We will assume the conditions on $\lambda_n$ stated in Theorems 2 and 3; that is, $\lambda_n / \sqrt{n} \to \lambda_0 \geq 0$ if $\gamma \geq 1$ and $\lambda_n / n^{\gamma/2} \to \lambda_0 \geq 0$ if $\gamma < 0$. The simple case is when all of the $\beta_j$'s are nonzero. Then $\hat{\beta}_{nj} \to_p \beta_j \neq 0$. Thus

$$V_n^*(\mathbf{u}) = -2\mathbf{u}^T \mathbf{W}_n^* + \mathbf{u}^T C_n \mathbf{u} + \lambda_0 \sum_{j=1}^{p} u_j \operatorname{sgn}(\beta_j) |\beta_j|^{\gamma - 1} + R_n^{(\gamma)}(\mathbf{u}) \quad \text{if } \gamma \geq 1$$

and

$$V_n^*(\mathbf{u}) = -2\mathbf{u}^T\mathbf{W}_n^* + \mathbf{u}^T C_n \mathbf{u} + R_n^{(\gamma)}(\mathbf{u}) \quad \text{if } \gamma < 1,$$

where $R_n^{(\gamma)}(\mathbf{u}) = o_p(1)$ for each $\mathbf{u}$ and $\mathbf{W}_n^*$ has a limiting Normal distribution (with covariance matrix $C$). From this it follows that

$$P^*\big(\sqrt{n}(\hat{\boldsymbol{\beta}}_n^* - \tilde{\boldsymbol{\beta}}_n) \in A\big) \to_p P(\operatorname{argmin}(V) \in A),$$

where $V$ is as defined in Theorems 2 or 3.

If one or more of $\beta_j$'s is 0 then the argument given above still works for $\gamma > 1$ but fails for $\gamma \le 1$; a more sophisticated argument is needed for this latter case. Suppose that $\sqrt{n}(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to \mathbf{U}$ a.s. Then under the conditions on $\lambda_n$ given above, we have for $\gamma = 1$,

$$V_n^*(\mathbf{u}) = -2\mathbf{u}^T\mathbf{W}_n^* + \mathbf{u}^T C_n \mathbf{u} + \lambda_0 \sum_{j=1}^{p} [u_j \operatorname{sgn}(\beta_j) I(\beta_j \neq 0) + (|u_j + \mathbf{U}_j| - |\mathbf{U}_j|) I(\beta_j = 0)] + \mathbf{R}_n^{(1)}(\mathbf{u})$$

and for $\gamma < 1$,

$$V_n^*(\mathbf{u}) = -2\mathbf{u}^T\mathbf{W}_n^* + \mathbf{u}^T C_n \mathbf{u} + \lambda_0 \sum_{j=1}^{p} (|u_j + \mathbf{U}_j|^\gamma - |\mathbf{U}_j|^\gamma) I(\beta_j = 0) + \mathbf{R}_n^{(\gamma)}(\mathbf{u}),$$

where $\mathbf{R}_n^{(\gamma)}(\mathbf{u}) = o(1)$ with probability 1. From this, it follows that

$$P^*\big(\sqrt{n}(\hat{\boldsymbol{\beta}}_n^* - \tilde{\boldsymbol{\beta}}_n) \in A\big) \to P^*(\operatorname{argmin}(V^*) \in A) \text{ a.s.},$$

where for $\gamma = 1$,

$$V^*(\mathbf{u}) = -2\mathbf{u}^T\mathbf{W}^* + \mathbf{u}^T C \mathbf{u} + \lambda_0 \sum_{j=1}^{p} [u_j \operatorname{sgn}(\beta_j) \mathbf{I}(\beta_j \neq 0) + |u_j + U_j| \mathbf{I}(\beta_j = 0)]$$

and for $\gamma < 1$,

$$V^*(\mathbf{u}) = -2\mathbf{u}^T\mathbf{W}^* + \mathbf{u}^T C \mathbf{u} + \lambda_0 \sum_{j=1}^{p} |u_j + U_j|^\gamma I(\beta_j = 0).$$

Note the parallels between these results and the results of Section 3.

In our case, we do not have almost sure convergence of $\sqrt{n}(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ but rather convergence in distribution; however, by the Skorokhod representation theorem [(cf.) van der Vaart and Wellner (1996)], given $U_n \to_d U$ there exists a probability space and random elements $\{U_n'\}$, $U'$ having the same distributions as $\{U_n\}$, $U$ such that $U_n' \to U'$ a.s. From this fact, we can deduce that

$$P^*\big(\sqrt{n}(\hat{\boldsymbol{\beta}}_n^* - \tilde{\boldsymbol{\beta}}_n) \in A\big) \to_d P^*(\operatorname{argmin}(V^*) \in A),$$

where probability in the limit is in fact a random variable if $\beta_j = 0$ for at least one $j$. On the other hand, if $\beta_j \neq 0$ for all $j$ then the limiting probability is nonrandom and is the same as that given in Theorems 2 and 3.

The asymptotic results presented above indicate that the bootstrap may have some problems in estimating the sampling distribution of Bridge estimators for $\gamma < 1$ when some true parameter values are either exactly 0 or close to 0; in such cases, bootstrap sampling introduces a bias (due to $\tilde{\boldsymbol{\beta}}_n$) that does not vanish asymptotically. One possible solution is to choose an estimator $\tilde{\boldsymbol{\beta}}_n$ that has $P(\tilde{\beta}_{nj} = 0) \approx 1$ when $\beta_j = 0$ but $P(\tilde{\beta}_{nj} = 0) \approx 0$ when $\beta_j \neq 0$; there are a variety of ways to do this, for example, by using a consistent model selection procedure. While this may seem attractive from an asymptotic viewpoint, such an approach may cause more problems in practice than it solves.

**5. Asymptotics for nearly singular designs.** In this section, we will consider the asymptotic behavior of Bridge estimators when the design is nearly singular. More precisely, suppose that $C_n$ [as defined in (3)] is nonsingular but tends to a singular matrix $C$. In particular, we will assume that

$$a_n(C_n - C) \to D \tag{15}$$

for some sequence $\{a_n\}$ tending to infinity where $D$ is positive definite on the null space of $C$ (that is, $\mathbf{v}^T D\mathbf{v} > 0$ for nonzero $\mathbf{v}$ with $C\mathbf{v} = \mathbf{0}$). (Note that $D$ is necessarily nonnegative definite on the null space of $C$ so that it is not too stringent to require it to be positive definite on this null space.)

The consistency and limiting distribution arguments given in Section 2 require that the functions $Z$ and $V$ (defined in Theorems 1, 2 and 3) have unique minimizers. When the matrix $C$ is singular this is not generally the case. For example, define $V(\mathbf{u})$ as in Theorem 2. If $\gamma > 1$ then $\mathbf{u} \in \operatorname{argmin}(V)$ satisfies

$$C\mathbf{u} - \lambda_0 \tau(\mathbf{W}, \boldsymbol{\beta}) = \mathbf{0} \tag{16}$$

for some function $\tau$. If $\mathbf{v}$ lies in the null space of $C$ then clearly $\mathbf{u} + t\mathbf{v} \in \operatorname{argmin}(V)$ for any $t$ and so $\operatorname{argmin}(V)$ consists of a single point if, and only if, $C$ is nonsingular. Likewise, when $\gamma = 1$, $\mathbf{u} \in \operatorname{argmin}(V)$ satisfies a modification of (16), namely

$$C\mathbf{u} - \lambda_0 \tau(\mathbf{W}, \boldsymbol{\beta}) \ni \mathbf{0}, \tag{17}$$

where now $\tau$ is possibly a set-valued function (or multifunction). Again $\mathbf{u} + t\mathbf{v} \in \operatorname{argmin}(V)$ for any $\mathbf{v}$ in the null space of $C$ and so (17) has a unique solution if, and only if, $C$ is nonsingular.

When $\gamma < 1$, the situation is somewhat more complicated. Define $V(\mathbf{u})$ as in Theorem 3. In general, if $C$ is singular then $\operatorname{argmin}(V)$ will not be unique; if $\mathbf{u} \in \operatorname{argmin}(V)$ and $\mathbf{v}$ lies in the null space of $C$ then for some nonzero $t$, $V(\mathbf{u}) = V(\mathbf{u} + t\mathbf{v})$. However, suppose that $\beta_{r+1} = \cdots = \beta_p = 0$ and that the null space of $C$ is spanned by the standard basis vectors $e_{r+1}, \ldots, e_p$; then we have

$$V(\mathbf{u}) = V_0(u_1, \ldots, u_r) + \lambda_0 \sum_{j=r+1}^{p} |u_j|^\gamma,$$

which has a unique minimizer. Note that this condition on the null space of $C$ implies that the strongest collinearity in the design is restricted to the covariates that have no influence on the response.

We will now consider the asymptotic behavior of nearly singular designs under fairly weak conditions. We will assume that $C_n$ is nonsingular for all $n$ and satisfies (15) for some sequence $\{a_n\}$. Define $b_n = (n/a_n)^{1/2}$ and redefine $V_n$ to be

$$V_n(\mathbf{u}) = \sum_{i=1}^{n}\left[(\varepsilon_i - \mathbf{u}^T\mathbf{x}_i/b_n)^2 - \varepsilon_i^2\right] + \lambda_n \sum_{j=1}^{p}\left[|\beta_j + u_j/b_n|^\gamma - |\beta_j|^\gamma\right] \tag{18}$$

Note that since $b_n = o(\sqrt{n})$, the estimators will have a slower rate of convergence than when $C$ is nonsingular.

THEOREM 6. *Assume a nearly singular model with $C_n$ satisfying* (15). *Let* $\mathbf{W}$ *be a zero mean multivariate Normal random vector such that* $\mathrm{Var}(\mathbf{u}^T\mathbf{W}) = \mathbf{u}^T D\mathbf{u} > 0$ *for each nonzero* $\mathbf{u}$ *satisfying* $C\mathbf{u} = \mathbf{0}$.

(a) *If* $\gamma > 1$ *and* $\lambda_n/b_n \to \lambda_0 \geq 0$, *then*

$$b_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to_d \operatorname{argmin}\{V(\mathbf{u})\colon C\mathbf{u} = \mathbf{0}\},$$

*where*

$$V(\mathbf{u}) = -2\mathbf{u}^T\mathbf{W} + \mathbf{u}^T D\mathbf{u} + \lambda_0 \sum_{j=1}^{p} u_j \operatorname{sgn}(\beta_j)|\beta_j|^{\gamma-1}.$$

(b) *If* $\gamma = 1$ *and* $\lambda_n/b_n \to \lambda_0 \geq 0$, *then*

$$b_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to_d \operatorname{argmin}\{V(\mathbf{u})\colon C\mathbf{u} = \mathbf{0}\},$$

*where*

$$V(\mathbf{u}) = -2\mathbf{u}^T\mathbf{W} + \mathbf{u}^T D\mathbf{u} + \lambda_0 \sum_{j=1}^{p}[u_j \operatorname{sgn}(\beta_j)\mathbf{I}(\beta \neq 0) + |u_j|\mathbf{I}(\beta_j = 0)].$$

(c) *If* $\gamma < 1$ *and* $\lambda_n/b_n^\gamma \to \lambda_0 \geq 0$ *then*

$$b_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to_d \operatorname{argmin}\{V(\mathbf{u})\colon C\mathbf{u} = \mathbf{0}\},$$

*where*

$$V(\mathbf{u}) = -2\mathbf{u}^T\mathbf{W} + \mathbf{u}^T D\mathbf{u} + \lambda_0 \sum_{j=1}^{p} |u_j|^\gamma I(\beta_j = 0).$$

PROOF. The proofs of (a), (b) and (c) are essentially the same as before. Define $V_n$ as in (18). Then in each case, for $u$ in the null space of $C$, we have $V_n(\mathbf{u}) \to_d V(\mathbf{u})$ while for $\mathbf{u}$ outside this null space, $V_n(\mathbf{u}) \to_p \infty$. □

EXAMPLE 2. Consider a design with

$$C_n = \begin{pmatrix} 1 & \rho_n & \cdots & \rho_n \\ \rho_n & 1 & \cdots & \rho_n \\ \vdots & \vdots & \ddots & \vdots \\ \rho_n & \cdots & \rho_n & 1 \end{pmatrix},$$

where $\rho_n \to 1$ and $a_n(1-\rho_n) \to \psi > 0$. In this case, $\{C_n\}$ converges to a matrix $C$ (of all 1's) and $a_n(C_n - C) \to D$ where

$$D = \begin{pmatrix} 0 & -\psi & \cdots & -\psi \\ -\psi & 0 & \ldots & -\psi \\ \vdots & \vdots & \ddots & \vdots \\ -\psi & \cdots & -\psi & 0 \end{pmatrix}.$$

If the matrices are $p \times p$ then the null space of $C$ is the space of vectors $\mathbf{u}$ with $u_1 + \cdots + u_p = 0$. For the sake of illustration, let's suppose that $\beta_1 \neq 0$, $\beta_2 = \cdots = \beta_p = 0$ and take $\gamma < 1$. Then the limiting objective function $V$ in Theorem 6 is

$$V(\mathbf{u}) = -2\mathbf{u}^T\mathbf{W} + \mathbf{u}^T D\mathbf{u} + \lambda_0[|u_2|^\gamma + \cdots + |u_p|^\gamma] \quad \text{for } u_1 + \cdots + u_p = 0,$$

where

$$\lambda_0 = \lim_{n\to\infty} \lambda_n \left(\frac{a_n}{n}\right)^{\gamma/2}.$$

The limiting distribution of $(n/a_n)^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ is simply $\widehat{\mathbf{U}} = \operatorname{argmin}\{V(\mathbf{u}) : u_1 + \cdots + u_p = 0\}$. Each component of $\widehat{\mathbf{U}}$ has positive probability mass at 0 and these components must sum to 0. Thus, if one uses Bridge estimation as a method for model selection (that is, estimate the number of nonzero parameters) then asymptotically the probability of selecting the true model is $P(\widehat{U} = \mathbf{0})$.

When $p = 2$ (with $\beta_1 \neq 0$ and $\beta_2 = 0$), it is relatively straightforward to compute the limiting distribution. In this case, define $u = u_1 = -u_2$ and $W = W_1 = -W_2$ (since $u_1 + u_2 = 0$ and $W_1 + W_2 = 0$) where $W \sim N(0, \psi)$. Then

$$V(u) = 2\psi u^2 - 4uW + \lambda_0|u|^\gamma.$$

It is possible to show (see Lemma A in the Appendix) that $V$ is minimized at 0 if

$$|W|^{2-\gamma} \leq \lambda_0 \psi^{1-\gamma} \left(\frac{2-\gamma}{2}\right)\left(\frac{2-\gamma}{2-2\gamma}\right)^{1-\gamma};$$

otherwise, $V$ is minimized at $\widehat{U}$ satisfying

$$\psi\widehat{U} + \frac{\lambda_0\gamma}{4}\frac{|\widehat{U}|^\gamma}{\widehat{U}} = W.$$

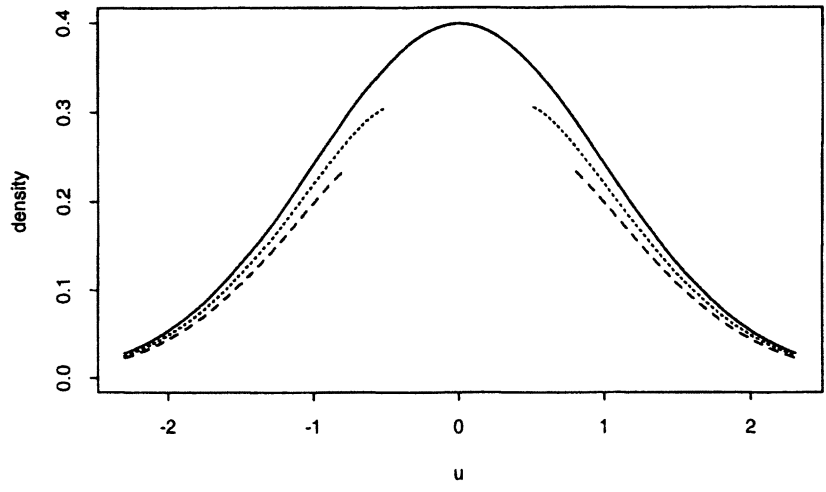


FIG. 4. *Densities for* $\lambda = 0$ *(solid line),* $\lambda = 0.5$ *(dotted line) and* $\lambda = 1$ *(dashed line); for* $\lambda = 0.5$ *and* $\lambda = 1$*; these are the densities of the absolute continuous part of the distribution as the distribution in these cases has positive probability mass at* 0.

The density of the absolutely continuous part of $\widehat{U}$ is

$$f(u) = \frac{4\psi - \lambda_0\gamma(1-\gamma)|u|^{\gamma-2}}{4\sqrt{2\pi\psi}} \exp\left[-\frac{1}{2\psi}\left(\psi u + \frac{\lambda_0\gamma}{4}\frac{|u|^\gamma}{u}\right)^2\right],$$

whenever

$$\left|\psi u + \frac{\lambda_0\gamma}{4}\frac{|u|^\gamma}{u}\right|^{2-\gamma} \geq \lambda_0\psi^{1-\gamma}\left(\frac{2-\gamma}{2}\right)\left(\frac{2-\gamma}{2-2\gamma}\right)^{1-\gamma} \tag{19}$$

and

$$|u|^{2-\gamma} \geq \frac{\lambda_0}{4\psi}\gamma(1-\gamma). \tag{20}$$

Setting $\psi = 1$ and $\gamma = 1/2$, the densities of the absolutely continuous part of $\widehat{U}$ for $\lambda_0 = 0, 0.5$ and 1 are given in Figure 4; for these parameter values $P(\widehat{U} = 0) = 0$, 0.448 and 0.655, respectively. Note that when $\lambda_0 > 0$, these densities have a "gap" [that is, $f(u) = 0$] for values of $u$ violating either or both of (19) and (20). □

## 6. Other issues.

*Singular designs.* In developing our asymptotic results, we have assumed almost exclusively in the previous sections that the matrix $C_n$ [defined in (3)] is nonsingular for each $n$ and hence that the parametrization in (1) is unique. In most situations, this is a reasonable assumption as a singular design can be made nonsingular by judiciously removing covariates or reparametrizing

the model. However, in some problems, singular designs are unavoidable. For example, in epidemiologic age-period-cohort studies of disease rates, singular designs result due to linear relationship among different variables [Kupper, Janis, Karmous and Greenberg (1985)]. Also in chemometric studies, singular designs result due to the number of parameters exceeding the number of observations [Frank and Friedman (1993)].

When $\lambda_n > 0$ and $\gamma > 1$, the objective function (2) is strictly convex and hence has a unique minimizer $\hat{\boldsymbol{\beta}}_n$; this may be true even for $\gamma \leq 1$. In this section, we will consider $n$ fixed and consider the behavior of the estimator as $\lambda = \lambda_n \to 0$.

Define $\hat{\boldsymbol{\beta}}_\lambda$ to minimize the objective function

$$\sum_{i=1}^{n}(Y_i - \mathbf{x}_i^T\boldsymbol{\phi})^2 + \lambda \sum_{j=1}^{p} |\phi_j|^\gamma \tag{21}$$

and note that if $\hat{\boldsymbol{\beta}}_\lambda$ minimizes (21) it also minimizes

$$h_\lambda(\boldsymbol{\phi}) = \frac{1}{\lambda}\left[\sum_{i=1}^{n}(Y_i - \mathbf{x}_i^T\boldsymbol{\phi})^2 - \sum_{i=1}^{n}\left(Y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}^{(0)}\right)^2\right] + \sum_{j=1}^{p} |\phi_j|^\gamma, \tag{22}$$

where $\hat{\boldsymbol{\beta}}^{(0)}$ is a LS estimator of $\boldsymbol{\beta}$, that is, $\hat{\boldsymbol{\beta}}^{(0)}$ satisfies

$$\sum_{i=1}^{n}\mathbf{x}_i\left(Y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}^{(0)}\right) = \mathbf{0}.$$

It is easy to see that as $\lambda \to 0$, $h_\lambda$ in (22) epiconverges to the function

$$h_0(\boldsymbol{\phi}) = \begin{cases} \sum_{j=1}^{p} |\phi_j|^\gamma, & \text{if } \sum_{i=1}^{n}\mathbf{x}_i(Y_i - \mathbf{x}_i^T\boldsymbol{\phi}) = \mathbf{0}, \\ \infty, & \text{otherwise} \end{cases} \tag{23}$$

and hence if $\operatorname{argmin}(h_0)$ is unique (as would be the case if $\gamma > 1$),

$$\hat{\boldsymbol{\beta}}_\lambda \to \hat{\boldsymbol{\beta}}_0 = \operatorname{argmin}\left\{\sum_{j=1}^{p} |\phi_j|^\gamma : \sum_{i=1}^{n}\mathbf{x}_i(Y_i - \mathbf{x}_i^T\boldsymbol{\phi}) = 0\right\} \tag{24}$$

as $\lambda \to 0$. When $\gamma = 2$, the estimator $\hat{\boldsymbol{\beta}}_0$ defined in (24) is simply a projection of the possible estimators onto the space spanned by the eigenvectors of $C_n$ with positive eigenvalues; this estimator is called the intrinsic estimator by Fu (1999). If we view $\hat{\boldsymbol{\beta}}_0$ as a regularized LS estimator then $\hat{\boldsymbol{\beta}}_\lambda$ can be used to approximate $\hat{\boldsymbol{\beta}}_0$ by taking $\lambda$ close to 0. Effectively, we are using an unconstrained optimization problem [minimizing $h_\lambda$ in (22) for small $\lambda$] to approximate a constrained optimization problem [minimizing $h_0$ in (23)]; this is a standard trick in optimization [Fiacco and McCormick (1990)].

*Computation.* To this point, we have not explicitly mentioned computation of the estimators. For $\gamma > 1$, the objective function is a smooth convex function and numerical algorithms such as Newton–Raphson or reweighted least squares work very well. For $\gamma = 1$, the objective function is also convex and so methods such as those discussed in Tibshirani (1996), Fu (1998) and Osborne, Presnell and Turlach (1998) can be used. In the context of wavelet regression, algorithms have been proposed by Chen, Donoho and Saunders (1999) and Sardy, Bruce and Tseng (1998).

When $\gamma < 1$, the objective function (2) is no longer convex and so computation of $\hat{\boldsymbol{\beta}}$ is nontrivial, particularly if $p$ is large; the objective function can have multiple local minima at which it is nondifferentiable. Here we will briefly describe some simple algorithms for computing Bridge estimates when $\gamma < 1$; a more detailed treatment will be given elsewhere.

Although the objective function is generally nontrivial to minimize, it is interesting to note that the one variable problem is quite easy to solve. For example, given $\alpha$ and $\lambda > 0$, define

$$g(u) = u^2 - 2\alpha u + \lambda |u|^\gamma.$$

It is simple to verify (see Lemma A in the Appendix) that $g$ is minimized at $u = 0$ if and only if

$$\lambda \geq \frac{2}{2-\gamma}\left(\frac{2-2\gamma}{2-\gamma}\right)^{1-\gamma} |\alpha|^{2-\gamma}.$$

Otherwise, $g$ is minimized at $u = \hat{u}$ satisfying $g'(\hat{u}) = 0$ and $g''(\hat{u}) > 0$. This latter equation can be solved in a variety of ways including the fixed-point iteration:

$$\hat{u}^{(0)} = \alpha,$$

$$\hat{u}^{(k)} = \alpha - \frac{\lambda\gamma}{2}\frac{|\hat{u}^{(k-1)}|^\gamma}{\hat{u}^{(k-1)}}, \qquad k = 1, 2, 3, \ldots .$$

The feasibility of the one-variable problem suggests that a Gauss–Seidel or ICM [Besag (1986)] algorithm (which iteratively minimize one variable at a time) might be appropriate to compute $\hat{\boldsymbol{\beta}}_n$. This is true to some extent (as the objective function decreases at each iteration) but with some caveats. Due to the nature of the objective function, it is very easy for a naive Gauss–Seidel algorithm to get "trapped" in a local minimum. However, this can be avoided to some extent by keeping estimates away from 0 until it is absolutely necessary to set them to 0. Alternatively, we can try multiple starting points in different parts of the parameter space.

A second approach is to solve a sequence of ridge regression problems. For example, starting with the ridge regression ($\gamma = 2$), estimate

$$\hat{\boldsymbol{\beta}}^{(0)} = \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T + \lambda I\right)^{-1}\left(\sum_{i=1}^n Y_i \mathbf{x}_i\right).$$

We can define successive estimates by

$$\hat{\boldsymbol{\beta}}^{(k)} = \Psi\big(\hat{\boldsymbol{\beta}}^{(k-1)}\big)\Big(\sum_{i=1}^{n} Y_i \mathbf{x}_i\Big).$$

for $k = 1, 2, 3, \ldots$ where

$$\Psi(\boldsymbol{\phi}) = D(\boldsymbol{\phi})\Big(D(\boldsymbol{\phi})\Big(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T\Big)D(\boldsymbol{\phi}) + \lambda I\Big)^{-1} D(\boldsymbol{\phi})$$

and $D(\phi)$ is a diagonal matrix with diagonal elements $|\phi_1|^{1-\gamma/2}, \ldots, |\phi_p|^{1-\gamma/2}$. Again the sequence $\{\hat{\boldsymbol{\beta}}^{(k)}\}$ does not necessarily converge to the global minimum but seems to work quite well if multiple starting points are used.

## APPENDIX

Let $g(u) = u^2 - 2\alpha u + \lambda|u|^\gamma$ where $\lambda \geq 0$ and $0 < \gamma \leq 1$. If $\alpha = 0$ then $\operatorname{argmin}(g) = 0$; thus we shall focus on the case where $\alpha \neq 0$.

LEMMA A. *Suppose that* $\alpha \neq 0$. *Then* $0 \in \operatorname{argmin}(g)$ *if, and only if,*

$$\lambda \geq |\alpha|^{2-\gamma}\Big(\frac{2}{2-\gamma}\Big)\Big(\frac{2(1-\gamma)}{2-\gamma}\Big)^{1-\gamma}.$$

*Moreover, if* $\gamma < 1$ *then* $\operatorname{argmin}(g) = 0$ *if and only if we have strict inequality above.*

PROOF. Define

$$h(t) = g(\alpha t) = \alpha^2(t^2 - 2t + \lambda|t|^\gamma|\alpha|^{\gamma-2}).$$

For $t < 0$, $h'(t) < 0$ and thus $h(t)$ is strictly decreasing on the interval $(-\infty, 0]$. For $t > 1$, $h'(t) > 0$ and so $h(t)$ is strictly increasing on the interval $(1, \infty)$. Thus $\operatorname{argmin}(g) = t\alpha$ for some $t \in [0, 1]$.

If $0 \in \operatorname{argmin}(g)$ then we must have

$$|\alpha|^{2-\gamma}(t^2 - 2t) + \lambda t^\gamma \geq 0$$

for all $0 \leq t \leq 1$. In other words,

$$\lambda \geq |\alpha|^{2-\gamma} \max_{0 \leq t \leq 1} t^{1-\gamma}(2 - t).$$

Using calculus, it is easy to verify that the right-hand side above is maximized for $t = 2(1-\gamma)/(2-\gamma)$ and so $0 \in \operatorname{argmin}(g)$ if and only if

$$\lambda \geq |\alpha|^{2-\gamma}\Big(\frac{2}{2-\gamma}\Big)\Big(\frac{2(1-\gamma)}{2-\gamma}\Big)^{1-\gamma}.$$

Moreover, if $\gamma < 1$ and $\alpha \neq 0$, then strict inequality implies that $\operatorname{argmin}(g) = 0$. If equality holds then $\operatorname{argmin}(g) = \{0, 2\alpha(1-\gamma)/(2-\gamma)\}$; note that this set contains a single point when $\gamma = 1$. □

**Acknowledgments.** The authors thank the referees and the Associate Editor for their valuable comments and suggestions. The support and encouragement of Rob Tibshirani is also gratefully acknowledged.

## REFERENCES

ANDERSON, P. K. and GILL, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10** 1100–1120.

BESAG, J. (1986). On the statistical analysis of dirty pictures (with discussion). *J. Roy. Statist. Soc. Ser. B* **48** 259–302.

CHEN, S. S., DONOHO, D. L. and SAUNDERS, M. A. (1999). Atomic decomposition by basis pursuit. *SIAM J. Scientific Computing* **20** 33–61.

FAN, J. and LI, R. (1999). Variable selection via penalized likelihood. Unpublished manuscript.

FIACCO, A. V. and MCCORMICK, G. P. (1990). *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. SIAM, Philadelphia.

FRANK, I. E. and FRIEDMAN, J. H. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics* **35** 109–148.

FU, W. J. (1998). Penalized regressions: the Bridge versus the Lasso. *J. Comput. Graph. Statist.* **7** 397–416.

FU, W. J. (1999). Estimating the effective trends in age-period-cohort studies. Unpublished manuscript.

GEYER, C. J. (1994). On the asymptotics of constrained $M$-estimation. *Ann. Statist.* **22** 1993–2010.

GEYER, C. J. (1996). On the asymptotics of convex stochastic optimization. Unpublished manuscript.

KIM, J. and POLLARD, D. (1990). Cube root asymptotics. *Ann. Statistics* **18** 191–219.

KUPPER, L. L., JANIS, J. M., KARMOUS, A. and GREENBERG, B. G. (1985). Statistical age-period-cohort analysis: a review and critique. *J. Chronic Disease* **38** 811–830.

LINHART, H. and ZUCCHINI, W. (1986). *Model Selection*. Wiley, New York.

OSBORNE, M. R., PRESNELL, B. and TURLACH, B. A. (1998). On the Lasso and its dual. Research Report 98/1, Dept. Statistics, Univ. Adelaide.

PFLUG, G. CH. (1995). Asymptotic stochastic programs. *Math. Oper. Res.* **20** 769–789.

POLLARD, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory* **7** 186–199.

SARDY, S., BRUCE, A. G. and TSENG, P. (1998). Block coordinate relaxation methods for nonparametric signal denoising with wavelet dictionaries. Technical Report, Dept. Mathematics, EPFL, Lausanne.

SRIVASTAVA, M. S. (1971). On fixed width confidence bounds for regression parameters. *Ann. Math. Statist.* **42** 1403–1411.

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

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer, New York.

DEPARTMENT OF STATISTICS
UNIVERSITY OF TORONTO
100 ST. GEORGE STREET
TORONTO, ONTARIO M5S 3G3
CANADA
E-MAIL: keith@utstat.toronto.edu

DEPARTMENT OF EPIDEMIOLOGY
MICHIGAN STATE UNIVERSITY
4660 S. HAGADORN RD., SUITE 600
EAST LANSING, MICHIGAN 48823
E-MAIL: fuw@msu.edu