Penalized Regressions: The Bridge Versus the Lasso

Wenjiang J. FU

Bridge regression, a special family of penalized regressions of a penalty function $\sum |\beta_j|^\gamma$ with $\gamma \geq 1$, is considered. A general approach to solve for the bridge estimator is developed. A new algorithm for the lasso ($\gamma = 1$) is obtained by studying the structure of the bridge estimators. The shrinkage parameter γ and the tuning parameter λ are selected via generalized cross-validation (GCV). Comparison between the bridge model ($\gamma \geq 1$) and several other shrinkage models, namely the ordinary least squares regression ($\lambda = 0$), the lasso ($\gamma = 1$) and ridge regression ($\gamma = 2$), is made through a simulation study. It is shown that the bridge regression performs well compared to the lasso and ridge regression. These methods are demonstrated through an analysis of a prostate cancer data. Some computational advantages and limitations are discussed.

Key Words: Bayesian prior; Bridge regressions; GCV; Newton–Raphson; Shrinkage; Shooting method.

1. INTRODUCTION

Consider a linear regression problem $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{y}$ is an n-vector of random responses, X an $n \times p$ design matrix, $\boldsymbol{\beta}$ a p-vector of parameters, and $\boldsymbol{\varepsilon}$ an n-vector of iid random errors. Ordinary least-squares regression (OLS), which minimizes RSS $= (\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta})$, yields an unbiased estimator $\hat{\boldsymbol{\beta}}_{\text{ols}} = (X^TX)^{-1}X^T\boldsymbol{y}$. Despite its simplicity and unbiasedness, the OLS estimator is, however, not always satisfactory because it is not unique if the design matrix X is less than full rank and the variance of the estimator $\text{var}(\hat{\boldsymbol{\beta}}_{\text{ols}}) = (X^TX)^{-1}\sigma^2$ is large if X is close to collinear. Therefore, mean squared error (MSE) is inflated by the collinearity and predictions based on $\hat{\boldsymbol{\beta}}_{\text{ols}}$ are overall not satisfactory. Detailed discussions can be found in Seber (1977), Sen and Srivastava (1990), Lawson and Hansen (1974), Hoerl and Kennard (1970a, 1970b) and Frank and Friedman (1993).

To achieve better prediction, Hoerl and Kennard (1970a, 1970b) introduced ridge regression, which minimizes RSS subject to a constraint $\sum |\beta_j|^2 \leq t$. Although ridge regression shrinks the OLS estimator towards 0 and yields a biased estimator $\hat{\boldsymbol{\beta}}_{\text{rdg}} = (X^TX + \lambda I)^{-1}X^T\boldsymbol{y}$, where $\lambda = \lambda(t)$, a function of t and I is an identity matrix, the

Wenjiang J. Fu is Assistant Professor, Department of Epidemiology, Michigan State University, East Lansing, MI 48823 (E-mail: fuw@pilot.msu.edu).

Journal of Computational and Graphical Statistics, Volume 7, Number 3, Pages 397–416

variance is smaller than that of the OLS estimator. Therefore, better estimation can be achieved on the average in terms of MSE with a little sacrifice of bias, and predictions can be improved overall. Frank and Friedman (1993) introduced bridge regression, which minimizes RSS subject to a constraint $\sum |\beta_j|^\gamma \le t$ with $\gamma \ge 0$. It includes ridge regression with $\gamma = 2$ and subset selection with $\gamma = 0$ as special cases. For different values of γ, the constrained areas are very different in the parameter space as shown in Figure 1 for $t = 1$. While Frank and Friedman (1993) did not solve for the estimator of bridge regression for any given $\gamma > 0$, they pointed out that it is desirable to optimize the parameter γ. Tibshirani (1996) introduced the lasso, which minimizes RSS subject to a constraint $\sum |\beta_j| \le t$, as a special case of the bridge with $\gamma = 1$. As pointed out by Tibshirani, the lasso shrinks the OLS estimator $\hat{\boldsymbol{\beta}}_{\text{ols}}$ towards 0 and potentially sets $\hat{\beta}_j = 0$ for some j. Thus, it performs as a variable selection operator. To solve for the lasso estimator, Tibshirani used a combined quadratic programming method by observing that the lasso constraint $\sum |\beta_j| \le t$ is equivalent to combining 2^p linear constraints $\sum w_j \beta_j \le t$ with $w_j = \pm 1$.

In this article, we study the structure of the bridge estimators and develop a general approach to solve bridge regression for $\gamma \ge 1$. Particularly we develop a simple algorithm for the lasso—the shooting method. The article is organized as follows. Section 2 gives the structure of the bridge estimators. The algorithms for the bridge and the lasso are developed in Section 3. The variance of the bridge estimator is derived in Section 4. The shrinkage parameter γ and the tuning parameter λ are selected for bridge regression via the generalized cross-validation (GCV) in Section 5. A special case of orthonormal matrix X is considered in Section 6. The bridge penalty is studied as a Bayesian prior in Section 7. The results of a simulation study are given in Section 8, and an analysis of a prostate cancer data is given in Section 9. Finally, some advantages of the shooting method for the lasso and some limitations of the model selection procedure via the GCV method are discussed. The mathematical proofs are given in the Appendix.

2. THE STRUCTURE OF THE BRIDGE ESTIMATORS

To solve bridge regression for any given $\gamma \ge 1$, we consider the following two problems.

$$\text{Given } \gamma \ge 1 \text{ and } t \ge 0, \quad \min_{\boldsymbol{\beta}} \text{RSS} \text{ subject to } \sum |\beta_j|^\gamma \le t. \tag{P1}$$

$$\text{Given } \gamma \ge 1 \text{ and } \lambda \ge 0, \quad \min_{\boldsymbol{\beta}} \left(\text{RSS} + \lambda \sum |\beta_j|^\gamma \right). \tag{P2}$$

Problems (P1) and (P2) are equivalent; that is, for given $0 \le \lambda \le +\infty$ there exists a $t \ge 0$, such that the two problems share the same solution, and vice versa. Problem (P1) is referred to as a constrained regression, while (P2) a penalized regression.

Consider problem (P2). Let $G(\boldsymbol{\beta}, X, \boldsymbol{y}, \lambda, \gamma) = \text{RSS} + \lambda \sum |\beta_j|^\gamma$. G is convex in $\boldsymbol{\beta}$, and $G \to +\infty$ as $||\boldsymbol{\beta}|| \to +\infty$. Thus, function G can be minimized; that is, there exists a $\hat{\boldsymbol{\beta}}$ such that $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} G(\boldsymbol{\beta}, X, \boldsymbol{y}, \lambda, \gamma)$. Take partial derivative of G with respect

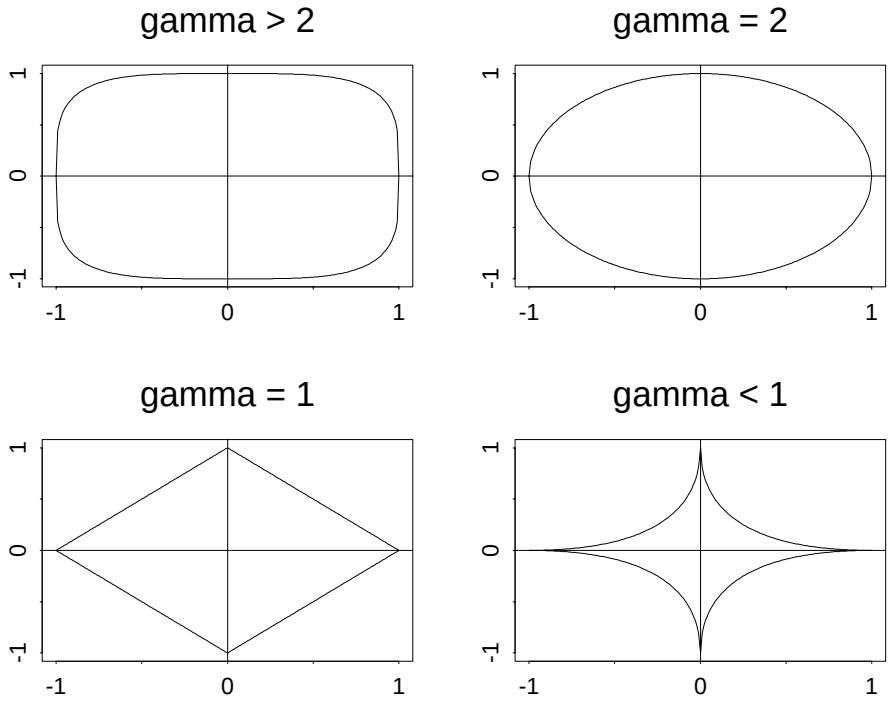

Figure 1. Constrained Areas of Bridge Regressions with t = 1.

to β_j at $\beta_j \neq 0$, $j = 1, \ldots, p$. Denote $S_j(\boldsymbol{\beta}, X, \boldsymbol{y}) = \partial \text{RSS}/\partial\beta_j$ and $d(\beta_j, \lambda, \gamma) = \lambda\gamma|\beta_j|^{\gamma-1}\text{sign}(\beta_j)$. Setting $\partial G/\partial\beta_j = 0$ leads to

$$\begin{cases} S_1(\boldsymbol{\beta}, X, \boldsymbol{y}) + d(\beta_1, \lambda, \gamma) &= 0 \\ \cdots & \\ S_p(\boldsymbol{\beta}, X, \boldsymbol{y}) + d(\beta_p, \lambda, \gamma) &= 0. \end{cases} \tag{P3}$$

Problem (P2) can then be solved through (P3). In fact, we have the following theorems on (P3) for more general function S_j.

Let $\boldsymbol{\beta}$ be a vector in a p-dimensional parameter space $\mathcal{B}$, X an $n \times p$ matrix, and $\boldsymbol{y}$ a vector in an n-dimensional sample space $\mathbf{R}^{\mathbf{n}}$. For fixed $X, \boldsymbol{y}, \lambda \geq 0, \gamma \geq 1$, define real functions $S_j(\cdot, X, \boldsymbol{y})$: $\mathcal{B} \longrightarrow \mathbf{R}, \boldsymbol{\beta} \longmapsto S_j(\boldsymbol{\beta}, X, \boldsymbol{y})$, $j = 1, \ldots, p$, and function $d(\beta_j, \lambda, \gamma) = \lambda\gamma|\beta_j|^{\gamma-1}\text{sign}(\beta_j)$. Denote $\mathbf{S} = (S_1, \ldots, S_p)^T$. We have the results for problem (P3).

Theorem 1. *Given $\gamma > 1$, $\lambda > 0$. If function $\mathbf{S}$ is continuously differentiable with respect to $\boldsymbol{\beta}$ and the Jacobian $(\partial \boldsymbol{S}/\partial\boldsymbol{\beta})$ is positive-semi-definite, then*

1. *(P3) has a unique solution $\hat{\boldsymbol{\beta}}(\lambda, \gamma)$, which is continuous in (λ, γ); and*
2. *the limit of the unique solution $\hat{\boldsymbol{\beta}}(\lambda, \gamma)$ exists as $\gamma \to 1+$. Denote the limit solution by $\hat{\boldsymbol{\beta}}(\lambda, 1+)$, then* $\lim_{\gamma \to 1+} \hat{\boldsymbol{\beta}}(\lambda, \gamma) = \hat{\boldsymbol{\beta}}(\lambda, 1+)$.

Theorem 2. *Given $\gamma > 1$, $\lambda > 0$. If functions S_j's are -2 multiples of the score functions of a joint likelihood function for Gaussian distribution, and the Jacobian $(\partial \boldsymbol{S}/\partial\boldsymbol{\beta})$ is positive definite, then*

1. *The unique solution of (P3) is equal to the unique estimator of the penalized regression (P2); and*
2. *the limit of the unique solution of (P3)* $\lim_{\gamma\to 1+}\hat{\boldsymbol{\beta}}(\lambda,\gamma)$ *is equal to the lasso estimator of (P2).*

The proofs are given in the appendix.

Remark 1. *Theorem 1 is independent of joint likelihood functions. Theorem 2 also holds for other distributions of the exponential family. Accordingly, problem (P2) must be modified by replacing RSS with the model deviance Dev*$(\boldsymbol{\beta}, X, \boldsymbol{y})$.

To solve bridge regression for any given $\gamma \geq 1$ and $\lambda > 0$, we start with problem (P3). Although we only demonstrate our method below for Gaussian distribution, our algorithms apply to other distributions in the exponential family via the iteratively reweighted least-squares (IRLS) procedure.

Denote $\boldsymbol{\beta}$ by $(\beta_j, \boldsymbol{\beta}^{-j})^T$, where $\boldsymbol{\beta}^{-j}$ is a $(p-1)$ vector consisting of the β_i's other than β_j. We study the jth equation of (P3):

$$S_j(\beta_j, \boldsymbol{\beta}^{-j}, X, \boldsymbol{y}) = -d(\beta_j, \lambda, \gamma). \tag{2.1}$$

The left hand side function of (2.1), LHS $= 2\boldsymbol{x}_j{}^T\boldsymbol{x}_j\beta_j + \sum_{i\neq j} 2\boldsymbol{x}_j{}^T\boldsymbol{x}_i\beta_i - 2\boldsymbol{x}_j{}^T\boldsymbol{y}$, is, for fixed $\boldsymbol{\beta}^{-j}$, a linear function of β_j with positive slope $2\boldsymbol{x}_j{}^T\boldsymbol{x}_j$. The right hand side function of (2.1), RHS $= -\lambda\gamma|\beta_j|^{\gamma-1}\text{sign}(\beta_j)$, is nonlinear in β_j. RHS is of different shape for different value of γ as shown in Figure 2. It is continuous, differentiable, and monotonically decreasing for $\gamma > 1$ except nondifferentiable at $\beta_j = 0$ for $1 < \gamma < 2$, a heavy-side function with a jump of height 2λ at $\beta_j = 0$ for $\gamma = 1$. Therefore, equation (2.1) has a unique solution for $\gamma > 1$, a unique solution or no solution for $\gamma = 1$.

3. ALGORITHMS FOR THE BRIDGE AND THE LASSO ESTIMATORS

To compute the bridge estimator for $\gamma > 1$, the Newton–Raphson method may apply. However, because function d is not differentiable at $\beta_j = 0$ for $\gamma < 2$, modification is needed to achieve the convergence to the solution. We develop the following modified Newton–Raphson method for $\gamma > 1$ in general by solving iteratively for the unique solution of the jth equation of (P3).

Modified Newton–Raphson (M-N-R) Algorithm for the Bridge $\gamma > 1$

(i) Start with $\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}}_{\text{ols}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$.

(ii) At step m, for each $j = 1, \ldots, p$, let $S_0 = S_j(0, \hat{\boldsymbol{\beta}}^{-j}, X, \boldsymbol{y})$. Set $\hat{\beta}_j = 0$ if $S_0 = 0$. Otherwise, if $\gamma \geq 2$, apply the Newton–Raphson method to solve for the unique solution $\hat{\beta}_j$ of equation (2.1); if $\gamma < 2$, modify function $-d$ by changing one part to its tangent line at some point between the solution and the origin as shown in Figure 3 (upper left figure). Then apply the Newton–Raphson method to equation (2.1) with the modified function $-d$ to solve for the unique solution $\hat{\beta}_j$. Form a new estimator $\hat{\boldsymbol{\beta}}_m = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ after updating all $\hat{\beta}_j$.

(iii) Repeat (ii) until $\hat{\boldsymbol{\beta}}_m$ converges.

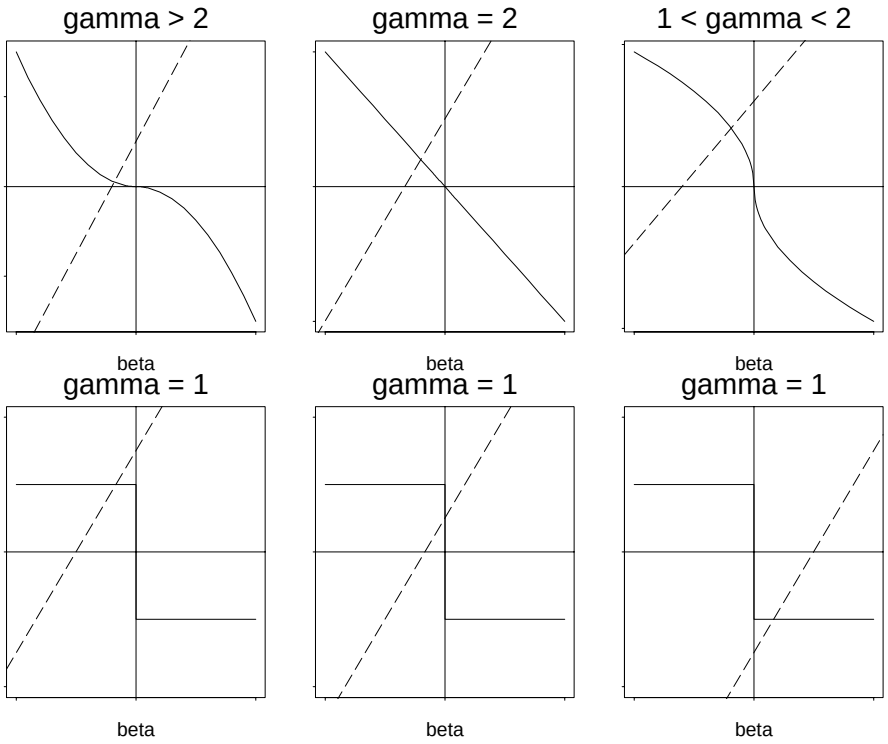

Figure 2. The Functions in Equation (2.1). Solid is function $-d$, dashed is S_j.

To compute the lasso estimator for any given $\lambda > 0$, one can apply the result of Theorem 1; that is, the limit of the bridge estimator, $\lim_{\gamma \to 1+} \hat{\boldsymbol{\beta}}(\lambda, \gamma)$, is equal to the lasso estimator. However, taking the limit numerically is not recommended in practice for the following reasons. From the computational point of view, it is obviously time-consuming because the modified Newton–Raphson algorithm has to be run for each $\gamma_{(i)}$ in a series of $\{\gamma_{(i)}\}$ with $\gamma_{(i)} \to 1+$. From the theoretical point of view, it is misleading. Assume a sequence of $\{\gamma_{(i)}\}$ tends to $1+$, and the corresponding estimators have one coordinate $\{\hat{\beta}_{(i)}\}$ with the values $.1, \ldots, 10^{-6}, \ldots$ Numerically one cannot determine whether the limit of $\hat{\beta}_{(i)}$ is equal to 0. However, taking the limit theoretically leads to a new algorithm for the lasso, which is simple, straightforward, and fast as shown in the following.

We introduce a new algorithm for the lasso—the shooting method.

1. $p = 1$. Start with an initial estimator $\hat{\beta}_0$, the OLS estimator. From the point $(\hat{\beta}_0, 0)$, shoot in the direction of slope $2\boldsymbol{x}^T\boldsymbol{x}$ as shown in Figure 3. If a point on the ceiling is hit (upper right figure), or a point on the floor is hit (lower right figure), equation (P3) has a unique solution $\hat{\beta}$, which has a simple close form and is equal to the lasso estimator. If no point is hit—that is, shooting through the window (lower left figure)—then equation (P3) has no solution. One can take the limit of the bridge estimator theoretically. It is easy to prove that $\lim_{\gamma \to 1+} \hat{\beta}(\lambda, \gamma) = 0$. Therefore, set $\hat{\beta} = 0$ for the lasso estimator.

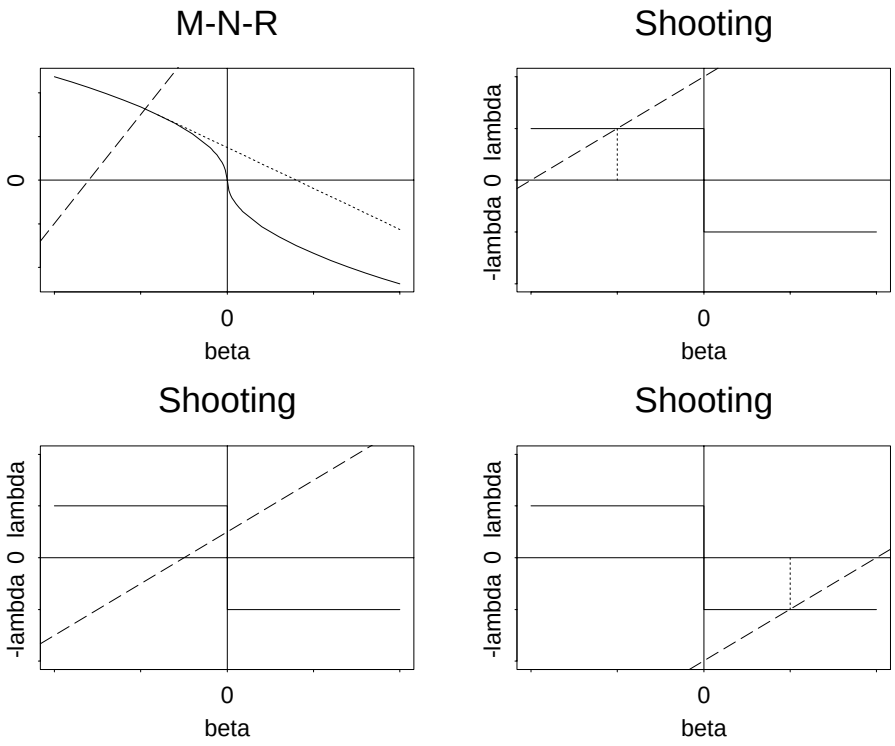

Figure 3. The Algorithms. Solid is function $-d$, dashed is S_j. Upper left: the dotted line represents the modification of $-d$ to its tangent; Upper right: $S_0 > \lambda$, the dotted line indicates the solution of (2.1); Lower left: $|S_0| \leq \lambda$; Lower right: $S_0 < -\lambda$, the dotted line indicates the solution of (2.1).

2. $p > 1$. Start with an initial value $\hat{\boldsymbol{\beta}}_0$. At step m, compute $\hat{\boldsymbol{\beta}}_m$ by updating $\hat{\beta}_j$ for fixed $\hat{\boldsymbol{\beta}}^{-j}$ using Step 1, $j = 1, \ldots, p$. Iterate until $\hat{\boldsymbol{\beta}}_m$ converges. We summarize the method as follows.

Shooting Algorithm for the Lasso

(i) Start with $\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}}_{\text{ols}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$.

(ii) At step m, for each $j = 1, \ldots, p$, let $S_0 = S_j(0, \hat{\boldsymbol{\beta}}^{-j}, X, \boldsymbol{y})$ and set

$$\hat{\beta}_j = \begin{cases} \frac{\lambda - S_0}{2\boldsymbol{x}_j{}^T \boldsymbol{x}_j} & \text{if} \quad S_0 > \lambda \\ \frac{-\lambda - S_0}{2\boldsymbol{x}_j{}^T \boldsymbol{x}_j} & \text{if} \quad S_0 < -\lambda \\ 0 & \text{if} \quad |S_0| \leq \lambda, \end{cases}$$

where $\boldsymbol{x}_j$ is the jth column vector of X. Form a new estimator $\hat{\boldsymbol{\beta}}_m = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ after updating all $\hat{\beta}_j$.

(iii) Repeat (ii) until $\hat{\boldsymbol{\beta}}_m$ converges.

Theorem 3. *(Convergence of the algorithms).*
Given fixed $\lambda > 0$ and $\gamma \geq 1$. $\hat{\boldsymbol{\beta}}_m$ in the modified Newton–Raphson algorithm

converges to the bridge estimator of (P2). $\hat{\boldsymbol{\beta}}_m$ in the shooting algorithm converges to the lasso estimator of (P2).

Our experience tells us that both the M-N-R and the shooting algorithms converge very fast, as it can be perceived through the mechanism of the convergence in the mathematical proof.

4. THE VARIANCE OF THE BRIDGE ESTIMATOR

Because the bridge estimator ($\gamma > 1$) is the unique solution of problem (P3) and is almost surely nonzero, its variance can be derived as follows from (P3) using the delta method.

$$\text{var}(\hat{\boldsymbol{\beta}}) = \left(X^T X + D(\hat{\boldsymbol{\beta}})\big|_{\boldsymbol{y}_0}\right)^{-1} X^T \text{var}(\boldsymbol{y}) X \left(X^T X + D(\hat{\boldsymbol{\beta}})\big|_{\boldsymbol{y}_0}\right)^{-1}, \qquad (4.1)$$

where $D(\hat{\boldsymbol{\beta}}) = \text{diag}\left(\lambda\gamma(\gamma-1)|\hat{\beta}_j|^{\gamma-2}/2\right)$ and $\boldsymbol{y}_0$ is an arbitrary fixed point in the sample space. The variance estimate can be obtained by plugging in $\hat{\boldsymbol{\beta}}$ for $\hat{\boldsymbol{\beta}}|_{\boldsymbol{y}_0}$ and replacing $\text{var}(\boldsymbol{y})$ with its estimate.

Denote $\boldsymbol{F} = (F_1, \ldots, F_p)^T$, where $F_j = S_j(\hat{\boldsymbol{\beta}}, X, \boldsymbol{y}) + d(\hat{\beta}_j, \lambda, \gamma)$. For Gaussian distribution, $\partial \boldsymbol{F}/\partial \boldsymbol{y} = -2X^T$ and $\partial \boldsymbol{F}/\partial \hat{\boldsymbol{\beta}} = 2X^T X + 2D(\hat{\boldsymbol{\beta}})$. Applying implicit function theorem to $\boldsymbol{F} = \mathbf{o}$, one has

$$\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{y}} = -\left(\frac{\partial \boldsymbol{F}}{\partial \hat{\boldsymbol{\beta}}}\right)^{-1} \frac{\partial \boldsymbol{F}}{\partial \boldsymbol{y}}.$$

Applying the delta method to the estimate $\hat{\boldsymbol{\beta}}$ as a function of $\boldsymbol{y}$ leads to the variance of $\hat{\boldsymbol{\beta}}$ in (4.1).

We examine this variance formula (4.1) in the following two special cases.

1. The OLS regression case—that is, $\lambda = 0$. The function $D(\hat{\boldsymbol{\beta}})$ becomes a zero matrix. Therefore $\text{var}(\hat{\boldsymbol{\beta}}) = \left(X^T X\right)^{-1} X^T \text{var}(\boldsymbol{y}) X \left(X^T X\right)^{-1}$, which equals to $\text{var}(\hat{\boldsymbol{\beta}}_{ols})$, the variance of the OLS estimator.
2. The ridge regression case—that is, $\gamma = 2$. The function $D(\hat{\boldsymbol{\beta}}) = \lambda I$, where I is an identity matrix. $\text{var}(\hat{\boldsymbol{\beta}}) = \left(X^T X + \lambda I\right)^{-1} X^T \text{var}(\boldsymbol{y}) X \left(X^T X + \lambda I\right)^{-1}$, which equals to $\text{var}(\hat{\boldsymbol{\beta}}_{\text{rdg}})$, the variance of the ridge estimator.

Since the lasso sets some $\hat{\beta}_j = 0$, the delta method does not apply. However, the bootstrap or the jackknife method can be used to compute the variance. A good variance estimator of the nonzero $\hat{\beta}_j$ of the lasso estimator can be found in Tibshirani (1996).

5. SELECTION OF THE SHRINKAGE PARAMETER γ AND THE TUNING PARAMETER λ

To select the parameters λ and γ, we use the generalized cross-validation (GCV) method (Craven and Wahba 1979), as suggested by Tibshirani (1996) for the lasso model as follows.

For given $\lambda \geq 0$ and $\gamma \geq 1$, compute the estimate $\hat{\boldsymbol{\beta}}$. The effective number of parameters of the model can then be computed as

$$p(\lambda) = \text{trace}\left(X(X^TX + \lambda W^-)^{-1}X^T\right) - n_0,$$

where W^- is the generalized inverse of $W = \text{diag}(2|\hat{\beta}_j|^{2-\gamma}/\gamma)$, and n_0 is the number of $\hat{\beta}_j$ such that $\hat{\beta}_j = 0$ for $\gamma = 1$, compensating the loss of inverse of entry zero on the diagonal of W due to $\hat{\beta}_j = 0$. This can be derived from (P3) as

$$\left(X^TX + \frac{\lambda\gamma}{2}\text{diag}(|\beta_j|^{\gamma-2})\right)\boldsymbol{\beta} = X^T\boldsymbol{y}.$$

Then define

$$\text{GCV} = \frac{\text{RSS}}{\text{n}(1 - \text{p}(\lambda)/\text{n})^2} \tag{5.1}$$

and select the values of λ and γ that minimize GCV over a grid of (λ, γ) as shown in Figure 9.

Remarks

1. For non-Gaussian distributions, deviance must be used to replace RSS in the GCV of (5.1).
2. The effective number of parameters defined here has an extra compensation term n_0 for the lasso ($\gamma = 1$) compared to the one in Tibshirani (1996). It also generalizes to accommodate for bridge regression with any $\gamma > 1$.

6. BRIDGE REGRESSIONS OF ORTHONORMAL MATRIX X

In this section, we study bridge regression of orthonormal regression matrix, a special case which allows us to study the characteristics of the shrinkage effect for different value of γ.

For orthonormal matrix $X = (x_{ij})$, $\sum_i x_{ij}x_{il} = 1$ if $j = l$, or 0 otherwise. Problem (P3) simplifies to p independent equations:

$$2\beta_j - 2\sum_i x_{ij}y_i + \lambda\gamma|\beta_j|^{\gamma-1}\text{sign}(\beta_j) = 0, \qquad j = 1, \ldots, p. \tag{6.1}$$

The estimator is then computed via the modified Newton–Raphson method for $\gamma > 1$ or via the shooting method for $\gamma = 1$. To study the shrinkage effect of different value of γ, we compare the bridge estimator—the solution of each single equation of (6.1), with the OLS estimator. Without making any confusion, we omit the subscript j of β_j and x_{ij} for simplicity.

Notice that equation (6.1) can be written as $\beta = \sum_i x_i y_i - \lambda\gamma|\beta|^{\gamma-1}\text{sign}(\beta)/2$. The first term on the right hand side is equal to the OLS estimator, the second term is due to the shrinkage and thus reflects the shrinkage effect. Therefore, $\hat{\beta}_{\text{brg}} = \hat{\beta}_{\text{ols}} - \lambda\gamma|\hat{\beta}_{\text{brg}}|^{\gamma-1}\text{sign}(\hat{\beta}_{\text{brg}})/2$.

To show the shrinkage effect of bridge regression, we plot the absolute value of the bridge estimator $\hat{\beta}_{\text{brg}}$, and compare it with the OLS estimator, whose absolute value is

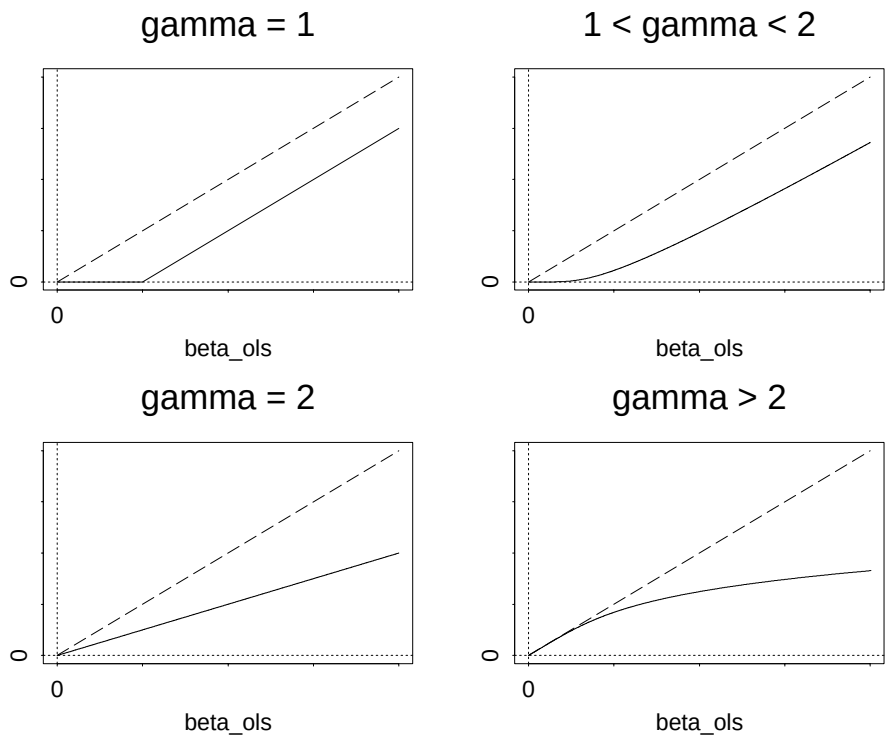

Figure 4. Shrinkage Effect of Bridge Regressions for Fixed $\lambda > 0$. Solid—the bridge estimator; dashed—the OLS estimator.

plotted on the diagonal as shown in Figure 4. It is shown that the lasso ($\gamma = 1$) shrinks small OLS estimates to zero and large by a constant; ridge regression ($\gamma = 2$) shrinks the OLS estimates proportionally; bridge regression ($1 < \gamma < 2$) shrinks small OLS estimates by a large rate and large by a small rate; bridge regression ($\gamma > 2$) shrinks small OLS estimates by a small rate and large by a large rate. In summary, bridge regression of large value of γ ($\gamma \geq 2$) tends to retain small parameters while small value of γ ($\gamma < 2$) tends to shrink small parameters to zero.

Therefore, it can be implied that if the true model includes many small but nonzero regression parameters, The lasso will perform poorly but the bridge of large γ value will perform well. If the true model includes many zero parameters, the lasso will perform well but the bridge of large γ value will perform poorly. Tibshirani (1996) obtained similar results by comparing the lasso with the ridge through intensive simulation studies.

7. BRIDGE PENALTY AS BAYESIAN PRIOR

In this section, we study the bridge penalty function $\sum |\beta_j|^\gamma$ as a Bayesian prior distribution of the parameter $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$. From the Bayesian point of view, bridge regression, $\min_{\boldsymbol{\beta}} \left(\text{RSS} + \lambda \sum |\beta_j|^\gamma\right)$, can be regarded as maximizing the log posterior

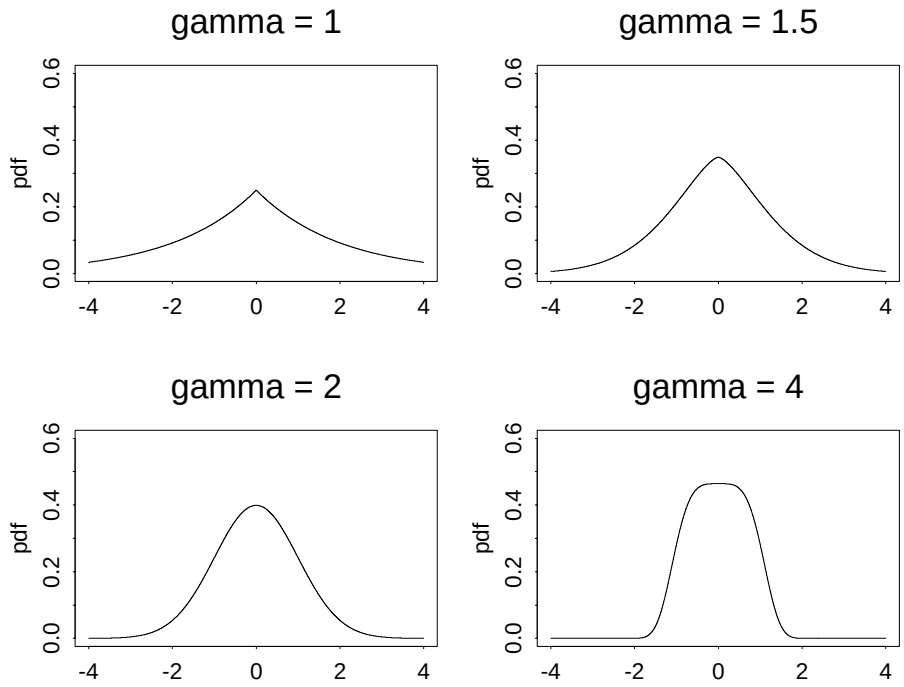

Figure 5. Bridge Penalty as a Bayesian Prior With λ =1.

distribution of

$$(\boldsymbol{\beta}|\boldsymbol{Y}) \sim C\exp\left\{-\frac{1}{2}\left(\text{RSS} + \lambda\sum|\beta_j|^\gamma\right)\right\},$$

where C is a constant. Thus, the bridge penalty $\lambda\sum|\beta_j|^\gamma$ can be regarded as the logarithm of the prior distribution $\exp\left\{-\frac{1}{2}\lambda\sum|\beta_j|^\gamma\right\}$ of the parameter $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ subject to a constant. Because the log prior is a summation, the parameters $\beta_1, \ldots, \beta_p$ are mutually independent and identically distributed. We thus omit the subscript j and study the prior distribution of β.

By simple algebra, the density function of the prior distribution of β is

$$\pi_{\lambda,\gamma}(\beta) = \frac{\gamma\, 2^{-(1+1/\gamma)}\lambda^{1/\gamma}}{\Gamma(1/\gamma)}\exp\left(-\frac{1}{2}\left|\frac{\beta}{\lambda^{-1/\gamma}}\right|^\gamma\right),$$

where $\lambda^{-1/\gamma}$ controls the window size of the density. Particularly, when $\gamma = 2$, β has a Gaussian distribution. Therefore, the posterior distribution of $(\boldsymbol{\beta}|\boldsymbol{Y})$ is also Gaussian if $\boldsymbol{Y}$ has a Gaussian distribution. This is a very special property of the ridge estimator for linear regressions.

To compare the penalty functions of different values of λ and γ, we plot the density function $\pi_{\lambda,\gamma}(\beta)$ in Figure 5 for $\lambda = 1$ and in Figure 6 for $\lambda = 10$. For $\lambda = 1$, it can be observed that small values of γ put much mass on the tails and the density has a large window and tends to be flat, while large values of γ put much mass in the center

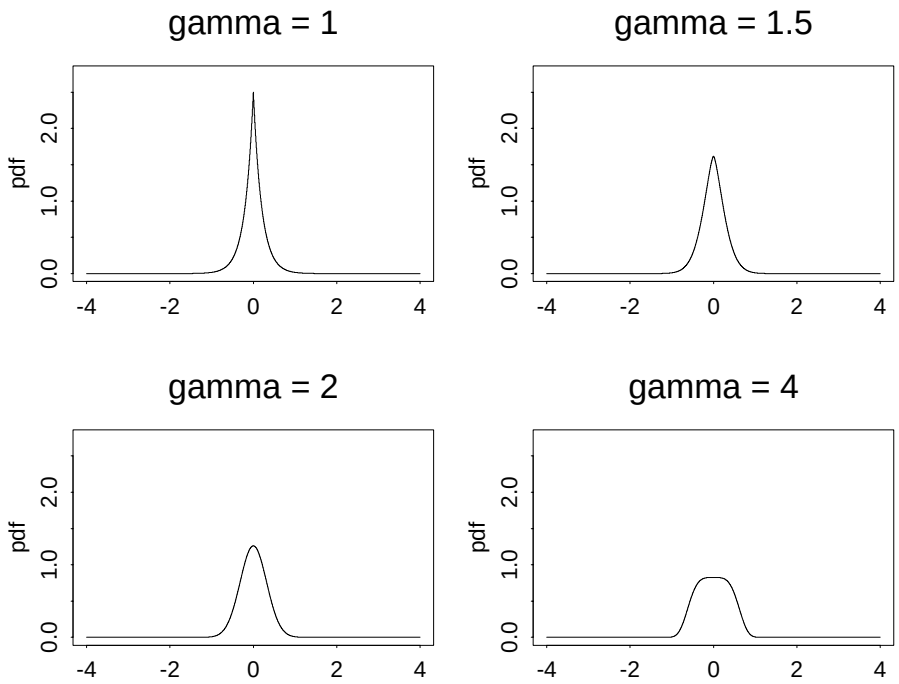

Figure 6. Bridge Penalty as a Bayesian Prior with λ =10.

around $\beta = 0$ and the density has a small window and is less spread out. For $\lambda = 10$, it can be observed that small values of γ put much mass close to 0 and the density has a very short tail, while large values of γ put mass relatively evenly in the window and the density tends to be uniform in the interval of $[-1,\ 1]$. When $\gamma = 2$, the density $\pi_{\lambda,\gamma}(\beta)$ is a Gaussian density.

It can thus be implied that the bridge penalty of small γ value favors models with regression parameters either of many zeros or of large absolute values from a long tailed density, while the bridge penalty of large γ value favors models with regression parameters of small but nonzero values from a normal-like or short tailed uniform-like density. Similar conclusion is reached by a simulation study in next section.

8. SIMULATION STUDY

We compare the bridge model with the OLS, the lasso and the ridge in a simulation of a linear regression model of 30 observations and 10 regressors

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{10} x_{10} + \varepsilon,$$

where $\varepsilon \sim N(0, \sigma^2)$. Ten regression matrices X_m, $m = 1, \ldots, 10$, are generated from an orthonormal matrix X of dimension 30×10 with different between-column pairwise correlation coefficients $\{\rho\}_m$ generated from uniform distribution $U(-1, 1)$.

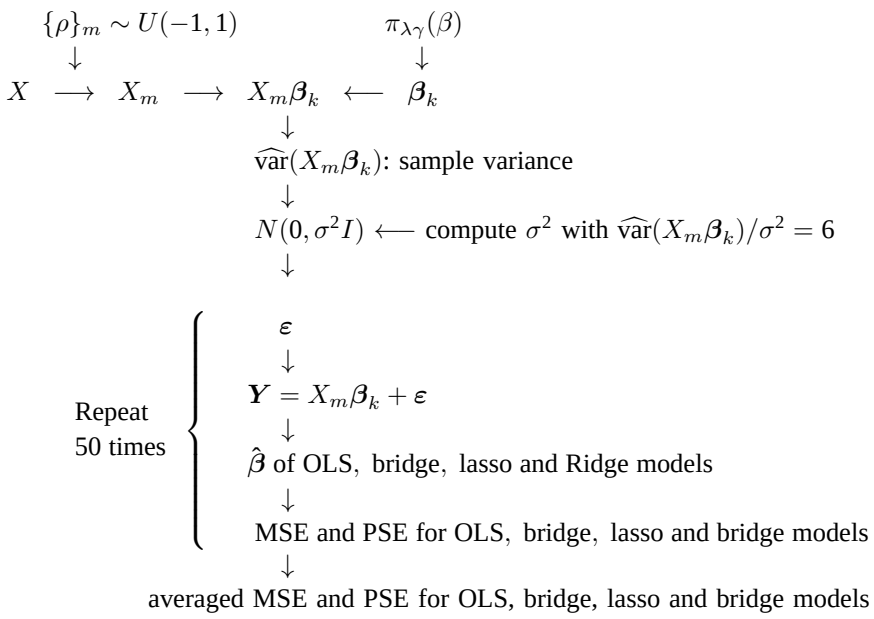

Figure 7. Schematic Diagram of the Data Generation for Fixed λ =1, $\gamma \geq$1, m = 1, . . ., 10 and k=1, . . ., 30.

The Data

For each X_m, 30 true $\boldsymbol{\beta}_k$, $k = 1, \ldots, 30$, are generated from the bridge prior $\pi_{\lambda\,\gamma}(\beta)$ with $\lambda = 1$ and fixed $\gamma \geq 1$. With each X_m and $\boldsymbol{\beta}_k$, 30 observations are generated from $\boldsymbol{Y} = X_m\boldsymbol{\beta}_k + \boldsymbol{\varepsilon}$ with iid normal random error ε_i from $N(0, \sigma^2)$ with a signal noise ratio $\widehat{\text{var}}(X_m\boldsymbol{\beta}_k)/\sigma^2 = 6$, where $v\hat{a}r$ is the sample variance of the vector. The MSE and PSE are computed as

$$\text{MSE} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T X_m^T X_m (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \quad \text{and} \quad \text{PSE} = (\boldsymbol{y} - X_m\hat{\boldsymbol{\beta}})^T (\boldsymbol{y} - X_m\hat{\boldsymbol{\beta}})$$

for different penalty models: the OLS, the bridge, the lasso, and the ridge. The PSE is computed as an average at 20 randomly selected points in the covariate space having the same correlation structure as X_m. Then the MSE and PSE are averaged over 50 replicates of the model with different random error $\boldsymbol{\varepsilon}$. Hence, for each $\boldsymbol{\beta}$ generated from the prior distribution $\pi_{\lambda,\gamma}(\beta)$, MSE and PSE are computed for the OLS, the bridge, the lasso, and the ridge models, as shown in the schematic diagram in Figure 7. Therefore $10 \times 30 = 300$ sets of MSE and PSE are computed for different $X_m\boldsymbol{\beta}_k$. The above procedure is repeated for different values of $\gamma = 1$, 1.5, 2, 3, 4.

The Method

Because each of the 300 sets of MSE and PSE is computed from different $X_m\boldsymbol{\beta}_k$ value, the variance σ^2 varies. The relationship PSE = MSE + σ^2 only holds within each set, but does not for the average over all sets. Hence we choose to compare the means of both MSE and PSE among different models. Because each set of MSE and PSE of different penalties are computed with the same $X_m\boldsymbol{\beta}_k$, and their values vary in a large

Table 1. Means* and SE's of MSE_r and PSE_r for different γ

	Bridge		*Lasso*		*Ridge*	
γ	MSE_r	PSE_r	MSE_r	PSE_r	MSE_r	PSE_r
1	.0860(.0044)	.0021(.0002)	.0841(.0043)	.0020(.0004)	.0595(.0030)	.0013(.0002)
1.5	.0225(.0054)	.0009(.0003)	.0224(.0054)	.0009(.0003)	.0566(.0032)	.0017(.0002)
2	–.0176(.0053)	.0002(.0005)	–.0176(.0053)	.0002(.0005)	.0519(.0028)	.0021(.0003)
3	–.0349(.0048)	–.0005(.0003)	–.0350(.0048)	–.0005(.0003)	.0566(.0029)	.0016(.0002)
4	–.0377(.0048)	-.0001(.0003)	–.0377(.0048)	–.0001(.0003)	.0577(.0027)	.0018(.0002)

* Minus sign means negative reduction; that is, increase of MSE or PSE.

range with different $X_m\boldsymbol{\beta}_k$ value, but the differences between the models are relatively small as shown in Figure 8, we compare the relative reduction of MSE and PSE from the OLS:

$$\text{MSE}_r = \frac{\text{MSE}_{\text{ols}} - \text{MSE}}{\text{MSE}_{\text{ols}}} \qquad \text{and} \qquad \text{PSE}_r = \frac{\text{PSE}_{\text{ols}} - \text{PSE}}{\text{PSE}_{\text{ols}}}.$$

It can be seen clearly from the plots of the MSE and PSE in the original scale that the MSE's of different penalty models are highly correlated, and so are the PSE's. It is appropriate to compare the relative reduction of MSE and PSE rather than the original MSE and PSE.

The Results

For each fixed γ value, the mean and its standard error of the 300 sets of MSE_r and PSE_r are computed and reported in Table 1. It is shown that for $\gamma = 1$ and $\gamma = 1.5$, the bridge, the lasso, and the ridge have significant reduction of MSE and PSE from the OLS. For $\gamma = 1$, the bridge has the greatest reduction, followed closely by the lasso, and by the ridge. For $\gamma = 1.5$, The ridge has the greatest reduction of MSE and PSE, followed by the bridge and by the lasso. For $\gamma = 2$, 3, and 4, the ridge has a significant reduction of MSE and PSE from the OLS, while the bridge and the lasso have a significant increase of MSE, and no significant reduction or increase of PSE.

It is demonstrated in Table 1 that the bridge has similar results to the lasso and performs well for small γ values, but not as well for large γ values. The ridge performs well for all of the γ values considered here. It performs better than the bridge and the lasso for large γ values but not as well for small γ values. As discussed in Section 7, large value of γ generates small but nonzero regression parameters from a short-tailed distribution, and small value of γ generates zeros or parameters of large absolute values from a long-tailed distribution. It can be implied that the bridge and the lasso may perform well if the true model has parameters of zeros or large absolute values from a long-tailed distribution, but perform poorly if the true model has many small but nonzero parameters from a short-tailed distribution. Such a result agrees with the results obtained in Sections 6 and 7. It also agrees with the results obtained by Tibshirani (1996) through intensive simulations.

In Figure 8, it shows on the right hand side the box plots of the MSE_r and PSE_r, and on the left hand side the plots of ten randomly selected sets of MSE and PSE in the original scale including the maximum and minimum. It is shown that the MSE's of different penalty models are highly correlated, and so are the PSE's. The values of the

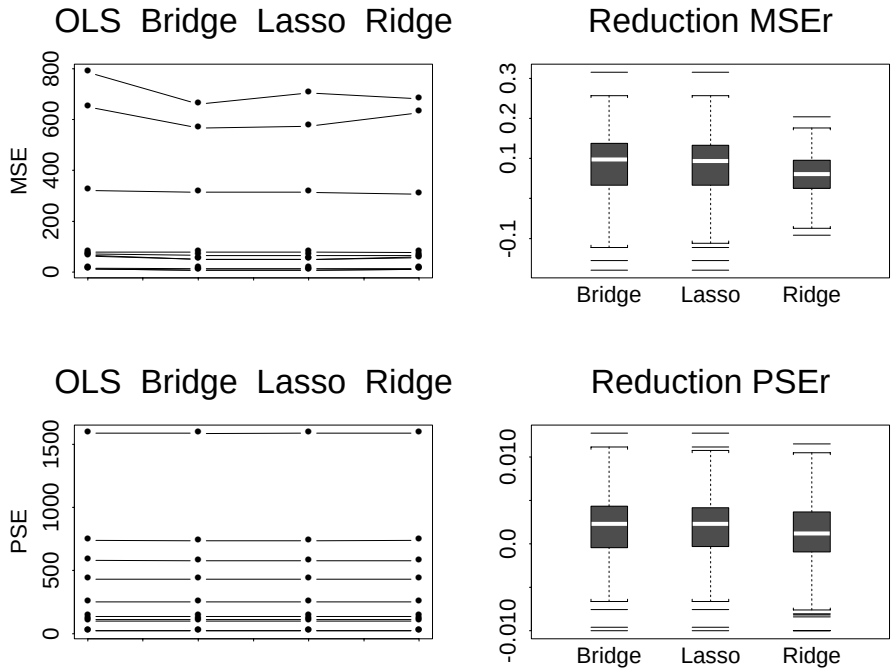

Figure 8. Comparison of MSE and PSE of Different Penalty Models by Simulation. MSE and PSE of different penalty models with true β generated from the bridge prior with $\gamma = 1$. Left: ten randomly selected sets of MSE and PSE including the maximum and minimum; Right: box plots of relative reductions from OLS: MSE_r and PSE_r.

MSE's and the PSE's vary in a large range. It can be inferred that the comparison of MSE_r and PSE_r between different penalty models is appropriate rather than the comparison of the original MSE and PSE.

Overall, bridge regression achieves small MSE and PSE, and performs well compared to the lasso and the ridge for linear regression models in general, but may perform poorly if the true models have many small but nonzero parameters.

9. ANALYSIS OF PROSTATE CANCER DATA

We apply the bridge penalty to a linear regression model to analyze a prostate cancer data. The data comes from a study by Stamey et al. (1989) that examined the correlation between the level of prostate specific antigen and a number of clinical measures in men who were about to receive a radical prostatectomy. The study had a total of 97 observations of male patients aged from 41 to 79 years. The covariates are log cancer volume (lcavol), log prostate weight (lweight), age of patient, log of benign prostatic hyperplasia amount (lbph), presence or absence of seminal vesicle invasion (svi), log of capsular penetration (lcp), Gleason grade (gleason), and percent Gleason grade 4 or 5 (pgg45). The data was later studied in Tibshirani (1996). A more detailed description of the data set can be found in either article.

We fit a linear model to the data. First, the data is centered by $\mathbf{x_j} = (\mathbf{x_j} - \bar{\mathbf{x}}_\mathbf{j})/\|\mathbf{x}_\mathrm{j} -$

Table 2. Correlation Matrix of the Covariates of the Prostate Cancer Data

lcavol	1.000	.194	.225	.027	.539	.675	.432	.434
lweight	.194	1.000	.308	.435	.109	.100	–.001	.051
age	.225	.308	1.000	.350	.118	.128	.269	.276
lbph	.027	.435	.350	1.000	–.086	-.007	.078	.078
svi	.539	.109	.118	–.086	1.000	.673	.320	.458
lcp	.675	.100	.128	–.007	.673	1.000	.515	.632
gleason	.432	–.001	.269	.078	.320	.515	1.000	.752
pgg45	.434	.051	.276	.078	.458	.632	.752	1.000

$\bar{\mathbf{x}}_\mathbf{j}\|$, where $\mathbf{x_j}$ is the jth column vector of the regression matrix X, and $\|\cdot\|$ is the Euclidean norm. Then a linear model is fitted to the centered data. Certain correlation is present between the covariates. For example, the pairwise coefficient is .752 between gleason and pgg45, .673 between svi and lcp, and .675 between lcavol and lcp, and so on. The condition number is 16.9, which indicates a slight collinearity in the covariates. All parameters of the OLS model are nonzero though some of them are not significant, for example, lcp, gleason, and pgg45. A bridge penalty model is also fitted and compared with the OLS estimator as in Table 3. For each pair of fixed $\lambda \geq 0$ and $\gamma \geq 1$, the bridge estimator is computed by either the M-N-R or the shooting algorithm. Then the GCV is computed as in (5.1). The penalty parameters are selected for this data set via the GCV as shown in Figure 9. A lasso model with $\lambda = 7.2$ is selected. The standard errors for the bridge estimates were computed by 10,000 bootstrap samples (Efron and Tibshirani 1993). Although the OLS model yields a significant effect of the intercept, lcavol, lweight, svi, and a marginal significant effect of age and lbph, the bridge model yields a significant effect of the intercept, lcavol, lweight, svi, and a marginal significant effect of lbph. The effect of age becomes nonsignificant in the bridge model. Two covariates—lcp and gleason—vanish in the bridge model.

We further compare the bridge model with the best model obtained from the subset selection by the leaps and bounds (L–B) method (Furnival and Wilson 1974; Seber 1977). The subset selection chooses the best model with the covariates lcavol, lweight, lbph, and svi. The covariates age and pgg45 are in the bridge model but not in the subset selection model. However, these two covariates are not significant at all based on their standard errors. Therefore, the bridge model agrees with the best model from the subset selection by the leaps and bounds method as shown in Table 4.

10. DISCUSSION

Bridge regression, as a special family of penalized regressions with two very important members—ridge regression and the lasso—plays an important role in solving collinearity problem. It yields small variance of the estimator and achieves good estimation and prediction by shrinking the estimator towards 0.

The simple and special structure of the bridge estimators for $\gamma \geq 1$ makes the computation very simple. The modified Newton–Raphson method for $\gamma > 1$ and the shooting method for $\gamma = 1$ were developed based on the theoretical results of the structure of

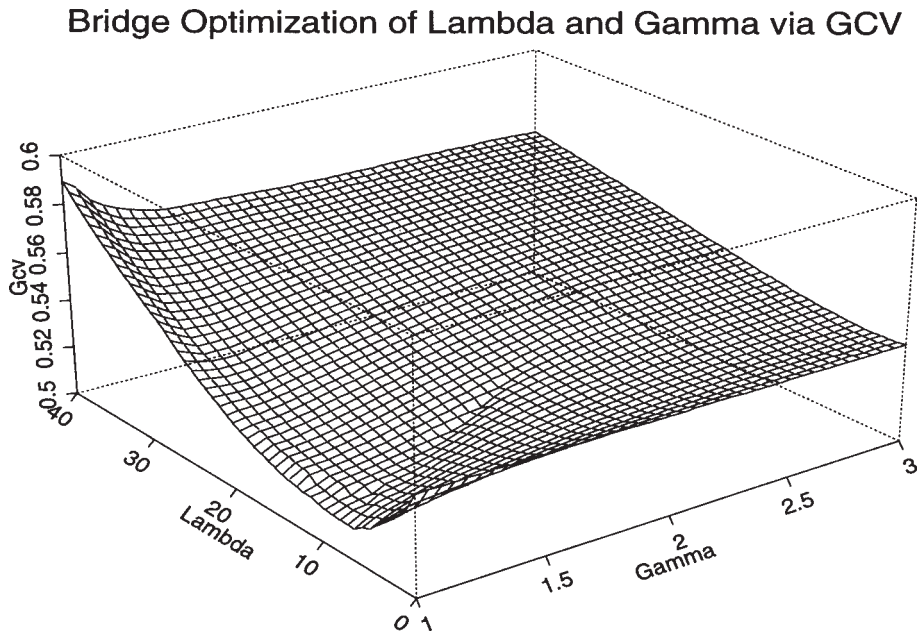

Figure 9. Selection of Parameters λ and γ for the Prostate Cancer Data.

the bridge estimators. Particularly, the shooting method for the lasso benefits from the theoretical result that the lasso estimator is the limit of the bridge estimator as γ tends to 1 from above. It has a very simple close form at each single step, and a simple iteration leads to fast convergence. These properties make it very attractive computationally in terms of CPU and memory. In contrast, the combined quadratic programming method by Tibshirani (1996) has a finite-step (2^p) convergence, and potentially has even better convergence rate ($.5p$ to $.75p$) as pointed out by Tibshirani (1996). In addition, the combined quadratic programming method has a standardized shrinkage rate $s = \sum_j |\hat{\beta}_{j\,\text{lasso}}|/|\hat{\beta}_{j\,\text{ols}}|$ as a tuning parameter, which has a range of $[0, 1]$ and is easy to optimize via grid search; while the shooting method has no such a standardized range, even though it has a thresh-

Table 3. The Estimates of the Prostate Cancer Data

	OLS	*Bridge*
intercept	2.478(.072)	2.478(.072)
lcavol	.688(.103)	.618(.090)
lweight	.225(.084)	.190(.076)
age	–.145(.082)	–.048(.046)
lbph	.155(.084)	.103(.066)
svi	.316(.100)	.245(.087)
lcp	–.147(.126)	.000(.068)
gleason	.032(.112)	.000(.047)
pgg45	.127(.123)	.063(.056)

Table 4. Comparison in Model Selection*

	OLS	*Bridge*	*Subset(L–B)*
lcavol	Y	Y	Y
lweight	Y	Y	Y
age	Y	N	N
lbph	Y	Y	Y
svi	Y	Y	Y
lcp	N	N	N
gleason	N	N	N
pgg45	N	N	N

* Y—significant effect; N— insignificant effect.

old $\lambda_0 > 0$ such that any tuning parameter $\lambda \geq \lambda_0$ sets the lasso estimates $\hat{\beta}_j = 0$ for $j = 1, \ldots, p$ (Gill, Murray, and Wright 1981). We believe that the shooting method has a convergence rate of order $p\log(p)$ although a theoretical result of the order has not been obtained. It is easy to see that for orthogonal X, only p steps is required to solve the p independent equations in (P3) by the shooting method. Both the modified Newton–Raphson method and the shooting method can be applied to generalized linear models via the IRLS procedure without extra effort.

The generalized cross-validation (GCV) method was proposed initially to optimize the tuning parameter of smoothing splines, which are linear operators. This technique is borrowed here to select the shrinkage parameters λ and γ, as suggested by Tibshirani (1996) for the lasso. It is evidently true in the literature that the GCV method works well for linear operators, including ridge regression. The simulation results of the linear regression model in Section 8 demonstrate that the GCV does not always select the best value of γ for bridge regression, even though bridge regression has the potential to select the best value of γ from a wide range $[1, \infty)$. The following fact may partially but not completely explain why the GCV does not select the best γ.

The bridge operator is nonlinear for $\gamma \neq 2$. This can be seen clearly from the equation in Section 5 due to the term $\text{diag}(|\beta_j|^{\gamma-2})$. The nonlinearity of the bridge operator can be visually seen in Figure 4 for the special case of orthonormal matrix. Since the bridge operator ($\gamma \neq 2$) performs very differently from the ridge operator ($\gamma = 2$) or the OLS operator ($\lambda = 0$), the linear approximation to the bridge operator as in the GCV definition (5.1) does not yield the best γ value for the model selection.

Because of the nonlinearity, it is not a surprise that the bridge model does not always perform the best in estimation and prediction compared to the other shrinkage models—the lasso and the ridge. Therefore, it is of great interest to investigate whether some other model selection methods, such as Mallow's C_p, AIC, or BIC criteria, perform well for bridge regression. If not, new optimization techniques are desirable, especially for nonlinear operators.

APPENDIX: MATHEMATICAL PROOFS

In this appendix, we give an outline of the mathematical proofs of Theorems 1, 2, and 3.

Denote $\boldsymbol{F} = (F_1, \ldots, F_p)^T$, where $F_j = S_j(\boldsymbol{\beta}, X, \boldsymbol{y}) + d(\beta_j, \lambda, \gamma)$, $j = 1, \ldots, p$. Equation (P3) is equivalent to $\boldsymbol{F} = \mathbf{o}$. We give two lemmas as follows.

Lemma 1. *Given $\lambda > 0$, $\gamma > 1$. If the Jacobian $\left(\partial \boldsymbol{S}/\partial \boldsymbol{\beta}\right)$ is positive-semi-definite, then $\left(\partial \boldsymbol{F}/\partial \boldsymbol{\beta}\right)$ is positive-definite at $\beta_j \neq 0$, $j = 1, \ldots, p$.*

Lemma 2. *Given $\lambda > 0$. Function $-d(\beta_j, \lambda, \gamma) = -\lambda\gamma|\beta_j|^{\gamma-1}sign(\beta_j)$ converges to the heavy-side function $-d(\beta_j, \lambda, 1) = -\lambda sign(\beta_j)$ at $\beta_j \neq 0$ as $\gamma \to 1+$.*

The proof of Lemma 1 is straight forward by observing that

$$\left(\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{\beta}}\right) = \left(\frac{\partial \boldsymbol{S}}{\partial \boldsymbol{\beta}}\right) + 2D(\boldsymbol{\beta}, \lambda, \gamma),$$

where $D(\boldsymbol{\beta}, \lambda, \gamma) = \text{diag}\left(\lambda\gamma(\gamma-1)|\beta_j|^{\gamma-2}/2\right)$. D is positive definite for $\gamma > 1$ and $\beta_j \neq 0$, $j = 1, \ldots, p$. The proof of Lemma 2 is obvious by observing that function d is continuous in γ at $\beta_j \neq 0$. Now we give a sketch of the proof of Theorem 1.

Proof of Theorem 1.

1. First, it is easy to prove the existence of the solution of (P3) and that the solution is almost surely nonzero by mathematical induction on dimension p. Second, the conditions of the implicit function theorem are satisfied by Lemma 1. Therefore, there exists a unique solution $\hat{\boldsymbol{\beta}}(\lambda, \gamma)$ satisfying (P3), and $\hat{\boldsymbol{\beta}}(\lambda, \gamma)$ is continuous in (λ, γ).
2. Now we prove the existence of the limit of $\hat{\boldsymbol{\beta}}(\lambda, \gamma)$ as $\gamma \to 1+$ by mathematical induction on dimension p.
   (a) $p = 1$. If there is an intersection of functions $S(\beta, X, \boldsymbol{y})$ and $-d(\beta, \lambda, 1)$ as shown in Figure 3 (upper right or lower right figure), by the continuity of functions $S(\beta, X, \boldsymbol{y})$ and $-d(\beta, \lambda, \gamma)$ and Lemma 2, the limit, $\lim_{\gamma \to 1+} \hat{\beta}(\lambda, \gamma)$, exists and is equal to the coordinate of the intersection. If there is no intersection of these two functions as shown in Figure 3 (lower left figure), it is easy to prove that $\lim_{\gamma \to 1+} \hat{\beta}(\lambda, \gamma) = 0$. Therefore, the result holds for $p = 1$.
   (b) For simplicity, we omit λ from the expressions since it is kept as a constant. Assume that the result holds for all dimensions $1, \ldots, (p-1)$. We prove that it also holds for dimension p. Consider a sub-problem formed by the first $p-1$ equations of (P3) for fixed β_p. By the assumption, the limit of the unique solution $(\hat{\beta}_1(\beta_p, \gamma), \ldots, \hat{\beta}_{p-1}(\beta_p, \gamma))$ of this sub-problem exists as $\gamma \to 1+$ for any fixed β_p. Plug into the last equation of (P3)

$$S_p\left(\hat{\beta}_1(\beta_p, \gamma), \ldots, \hat{\beta}_{p-1}(\beta_p, \gamma), \beta_p, X, \boldsymbol{y}\right) + d(\beta_p, \gamma) = 0. \tag{A.1}$$

   We prove that (A.1) has a unique solution $\hat{\beta}_p(\gamma)$ of which the limit exists as $\gamma \to 1+$. Denote the first term of the left hand side function of (A.1) by $L(\beta_p, \gamma)$. It can be proved that $\partial L/\partial \beta_p \geq 0$ by chain rule since the partial derivatives, $\partial \hat{\beta}_1/\partial \beta_p, \ldots, \partial \hat{\beta}_{p-1}/\partial \beta_p$ satisfy

$$\frac{\partial S_j}{\partial \hat{\beta}_1}\frac{\partial \hat{\beta}_1}{\partial \beta_p} + \cdots + \frac{\partial S_j}{\partial \hat{\beta}_{p-1}}\frac{\partial \hat{\beta}_{p-1}}{\partial \beta_p} + \frac{\partial S_j}{\partial \beta_p} = 0, \quad j = 1, \ldots, p-1 \tag{A.2}$$

   by the implicit function theorem on the subproblem. This implies the existence of the unique solution $\hat{\beta}_p(\gamma)$. Notice that $\partial L/\partial \beta_p \geq 0$ for any $\gamma > 1$. Similarly,

one can prove that the solution of the following equation exists.

$$S_p\left(\hat{\beta}_1(\beta_p, 1+), \ldots, \hat{\beta}_{p-1}(\beta_p, 1+), \beta_p, X, \boldsymbol{y}\right) + d(\beta_p, \gamma) = 0, \tag{A.3}$$

where $\hat{\beta}_j(\beta_p, 1+)$ is the limit of the solution $\hat{\beta}_j(\beta_p, \gamma)$ for fixed β_p, $j = 1, \ldots, p-1$. Denote the solution of (A.3) by $\tilde{\beta}_p(\gamma)$. Then $\lim_{\gamma \to 1+} \tilde{\beta}_p(\gamma)$ exists by the assumption of induction. Rewrite equation (A.1) as

$$S_p\left(\hat{\beta}_1(\beta_p, 1+), \ldots, \hat{\beta}_{p-1}(\beta_p, 1+), \beta_p, X, \boldsymbol{y}\right) + d(\beta_p, \gamma) + \Delta(\beta_p, \gamma) = 0, \tag{A.4}$$

where

$$\Delta(\beta_p, \gamma) = S_p\left(\hat{\beta}_1(\beta_p, \gamma), \ldots, \hat{\beta}_{p-1}(\beta_p, \gamma), \beta_p, X, \boldsymbol{y}\right) - S_p\left(\hat{\beta}_1(\beta_p, 1+), \ldots, \hat{\beta}_{p-1}(\beta_p, 1+), \beta_p, X, \boldsymbol{y}\right).$$

To prove that the limit of $\hat{\beta}_p(\gamma)$ exists, it suffices to prove that the solutions of (A.4) and (A.3) have the same limit. This can be achieved by the following inequality

$$|\Delta(\beta_p, \gamma)| \leq \delta(\gamma),$$

where $\delta(\gamma)$ is independent of β_p and converges to 0 as $\gamma \to 1+$. This inequality can be easily proved in functional analysis by observing that S_p is differentiable with bounded partial derivatives $\partial S_p / \partial \hat{\beta}_j$, $\hat{\beta}_j(\beta_p, \gamma)$ is differentiable with bounded partial derivatives $\partial \hat{\beta}_j / \partial \beta_p$ by Implicit Function Theorem, and $\hat{\beta}_j(\beta_p, \gamma) \to \hat{\beta}_j(\beta_p, 1+)$ for any value of β_p. This completes the proof of Theorem 1. □

Proof of Theorem 2.

1. Given $\lambda > 0$, $\gamma > 1$. Because there exists a joint likelihood function and $\partial \boldsymbol{S} / \partial \boldsymbol{\beta}$ is positive definite, the deviance function $-2\log(\text{Lik})$ is convex in $\boldsymbol{\beta}$. By the same argument of Lemma 1, function $G(\boldsymbol{\beta}, \lambda, \gamma) = -2\log(\text{Lik}) + \lambda \sum |\beta_j|^\gamma$ is convex and can be minimized uniquely at some finite point. Hence, the bridge estimator is unique. Because (P3) has a unique solution $\hat{\boldsymbol{\beta}}(\lambda, \gamma)$, which satisfies that $\hat{\beta}_j \neq 0$ almost surely for $j = 1, \ldots, p$, and function G is differentiable at $\hat{\boldsymbol{\beta}}(\lambda, \gamma)$, thus G attains the minimum at $\hat{\boldsymbol{\beta}}(\lambda, \gamma)$. By the uniqueness of the bridge estimator, $\hat{\boldsymbol{\beta}}(\lambda, \gamma)$ is equal to the bridge estimator of (P2).

2. Given $\lambda > 0$. By Theorem 1, $\lim_{\gamma \to 1+} \hat{\boldsymbol{\beta}}(\lambda, \gamma)$ exists. Denote the limit by $\hat{\boldsymbol{\beta}}(\lambda, 1+)$. Then $\lim_{\gamma \to 1+} G\left(\hat{\boldsymbol{\beta}}(\lambda, \gamma), \lambda, \gamma\right) = G\left(\hat{\boldsymbol{\beta}}(\lambda, 1+), \lambda, 1\right)$. Notice that $\hat{\boldsymbol{\beta}}(\lambda, \gamma)$ is the unique estimator minimizing $G(\lambda, \gamma)$, and $\hat{\boldsymbol{\beta}}_{\text{lasso}}$ is the unique estimator minimizing $G(\lambda, 1)$ since G is convex for $\gamma = 1$. It can be easily proved that $\hat{\boldsymbol{\beta}}(\lambda, 1+) = \hat{\boldsymbol{\beta}}_{\text{lasso}}$ by contradiction. □

Proof of Theorem 3. Because the limit of the bridge estimator is the lasso estimator as γ tends to $1+$, taking this limit at each step of the modified Newton-Raphson (M-N-R) algorithm leads to the shooting algorithm. Hence the convergence of the M-N-R algorithm implies the convergence of the shooting algorithm. Therefore it suffices to

prove the convergence of the M-N-R algorithm. We prove it for the case in which there exists a joint likelihood function.

Because there exists a joint likelihood function, function $G(\boldsymbol{\beta}, \lambda, \gamma)$ is convex by Lemma 1. Hence there exists a unique solution minimizing G—that is, $\hat{\boldsymbol{\beta}}_{\text{brg}} = arg\min G$. For $p = 1$, the M-N-R algorithm converges to the unique solution of (P3), which is the bridge estimator by Theorem 2. Hence, $\hat{\boldsymbol{\beta}}_{\text{brg}}$ minimizes function G. For $p > 1$ and fixed $\hat{\boldsymbol{\beta}}^{-j}$, updating $\hat{\beta}_j$ by M-N-R algorithm attains the local minimum of G in β_j for fixed $\hat{\boldsymbol{\beta}}^{-j}$. Denote the value of G by G_{mj} and the updated value of $\hat{\boldsymbol{\beta}}$ by $\hat{\boldsymbol{\beta}}_{mj}$ after updating $\hat{\beta}_j$ at step m by M-N-R algorithm, one has

$$G_{11} \geq G_{12} \geq \cdots \geq G_{m1} \geq \cdots \geq G_{mp} \geq \cdots \geq \min(G).$$

By the convexity of function G and the uniqueness of the bridge estimator, which minimizes G, G_{mj} converges to $\min(G)$, and $\hat{\boldsymbol{\beta}}_{mj}$ converges to $\hat{\boldsymbol{\beta}}_{\text{brg}}$, the unique bridge estimator. Consequently, the subsequence $\hat{\boldsymbol{\beta}}_{mp}$, which is the sequence $\hat{\boldsymbol{\beta}}_m$ in the M-N-R algorithm by definition, converges to the unique bridge estimator. □

ACKNOWLEDGMENTS

I am grateful to my PhD supervisor, R. Tibshirani, for introducing this interesting topic to me and for his supervision and advice. I also thank R. Neal and J. Hsieh for their helpful discussions and I. Johnston for access of the prostate cancer data. I appreciate the comments and suggestions of the associate editor, which greatly improved the appearance of the article.

[Received December 1997. Revised April 1998.]

REFERENCES

Craven, P., and Wahba, G. (1979), "Smoothing Noisy Data With Spline Functions," *Numerische Mathematik*, 31, 377–403.

Efron, B., and Tibshirani, R.J. (1993), *An Introduction to the Bootstrap*, New York: Chapman and Hall.

Frank, I.E., and Friedman, J.H. (1993), "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, 35, 109–148.

Furnival, G.M., and Wilson, R.W., Jr. (1974), "Regressions by Leaps and Bounds," *Technometrics*, 16, 499–511.

Gill, P.E., Murray, W., and Wright, M.H. (1981), *Practical Optimization*, London: Academic Press.

Hoerl, A.E., and Kennard, R.W. (1970a), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55–67.

——— (1970b), "Ridge Regression: Applications to Nonorthogonal Problems," *Technometrics*, 12, 69–82.

Lawson, C., and Hansen, R. (1974), *Solving Least Squares Problems*, Englewood Cliffs, NJ: Prentice-Hall.

Seber, G.A.F. (1977), *Linear Regression Analysis*, New York: Wiley.

Sen, A., and Srivastava, M. (1990), *Regression Analysis Theory, Methods, and Applications*, New York: Springer.

Stamey, T., Kabalin, J., McNeal, J., Johnston, I., Freiha, F., Redwine, E., and Yang, N. (1989), "Prostate Specific Antigen in the Diagnosis and Treatment of Adenocarcinoma of the Prostate ii. Radical Prostatectomy Treated Patients," *Journal of Urology*, 16, 1076–1083.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of Royal Statistical Society*, B, 58, 267–288.