L^γ Penalty Models
Computation And Applications
Part I

Wenjiang Fu

wfu@stat.tamu.edu

http://stat.tamu.edu/$\sim$wfu.

Department of Statistics, Texas A&M University

Outline

Motivation

Lecture 1. L^{γ} **penalty: variable selection and computation for linear models**

Lecture 2. **Selection of tuning parameter and asymptotics**

Lecture 3. **Extension to non-Gaussian response and longitudinal studies**

Lecture 4. **Recent development in** L^{γ} **penalty models and related topics**

Linear regression model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{Y}$ is $\boldsymbol{n}$-vector of responses, $\boldsymbol{\beta}$ is $\boldsymbol{p}$-vector of parameters, $\boldsymbol{X} = (\boldsymbol{x}_1 \ \ldots \ \boldsymbol{x}_p)$ is $\boldsymbol{n} \times \boldsymbol{p}$ matrix with column vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$, and $\boldsymbol{\varepsilon}$ is $\boldsymbol{n}$-vector of random errors with $\mathbf{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\mathbf{var}(\boldsymbol{\varepsilon}) = \boldsymbol{\sigma}^2 \boldsymbol{I}$.

Least-squares (LS) estimator $\widehat{\boldsymbol{\beta}}_{\mathbf{ols}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$, if X is of full rank, is BLUE (best linear unbiased estimator) .

$\mathbf{var}(\widehat{\boldsymbol{\beta}}_{\mathbf{ols}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{\sigma}^2$.

If column vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ are close to (but not exactly) linearly dependent, the vectors are said to be collinear. The determinant $\mathbf{det}(\boldsymbol{X}^T\boldsymbol{X})$ is close to 0. Then $\mathbf{var}(\widehat{\boldsymbol{\beta}}) \uparrow$.

Problems of LS estimator $\widehat{\beta}$ with collinearity

- Large variance and mean squared error.

$$\mathbf{MSE} = \mathbf{bias}^2 + \mathbf{var}.$$

- Poor estimation and prediction.

- Three major phenomena (Land *et al.*1990, AJS):

◇ Large changes in parameter estimate when adding or deleting variables;

◇ Wide confidence interval, nonsignificant test statistics, and opposite signs to expected values of important independent variables;

◇ Unstable regression parameters from sample to sample.

Diagnosis: condition number

Let $\lambda_1 \leq \ldots \leq \lambda_p$ be ordered eigenvalues of matrix $X^T X$.
The condition number is defined as $\sqrt{\lambda_p/\lambda_1}$. Cutoff: 30.

Q: How to improve performance?

James – Stein estimator.

If $\widehat{\theta} = x$ is an unbiased estimator for θ and $p \geq 3$, then
$J_x = \left(1 - \frac{p-2}{\|x\|_2^2}\right) x$ is called James – Stein estimator.

Shrinkage estimators.

Idea: Shrink parameters towards the origin
to reduce variance (bias-variance trade-off).

Recall: $\mathrm{MSE} = \mathrm{bias}^2 + \mathrm{var}$.

Ridge estimator (Hoerl and Kennard 1971)

$$\widehat{\boldsymbol{\beta}}_{\mathbf{rdg}} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} + \boldsymbol{\lambda I})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y},$$

where I is identity matrix, $\lambda \geq 0$ is tuning parameter.

$$\widehat{\boldsymbol{\beta}}_{\mathbf{rdg}} = \arg\min_{\boldsymbol{\beta}}\{(\boldsymbol{y} - \boldsymbol{X\beta})^{\boldsymbol{T}}(\boldsymbol{y} - \boldsymbol{X\beta}) + \boldsymbol{\lambda\beta^T\beta}\}.$$

Equivalently,
$\widehat{\boldsymbol{\beta}}_{\mathbf{rdg}} = \arg\min_{\boldsymbol{\beta}}\{(\boldsymbol{y} - \boldsymbol{X\beta})^{\boldsymbol{T}}(\boldsymbol{y} - \boldsymbol{X\beta})\}$ subject to $\boldsymbol{\beta^T\beta} \leq \boldsymbol{t}$,
with $t \geq 0$.

$$\mathbf{var}(\widehat{\boldsymbol{\beta}}_{\mathbf{rdg}}) \leq \mathbf{var}(\widehat{\boldsymbol{\beta}}_{\mathbf{ols}})$$

L^γ PENALTY

Bridge estimator (Frank and Friedman 1993)

$\widehat{\beta}_{\mathbf{brdg}} = \arg\min_{\beta}\{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{\lambda}\sum_{j=1}^{p} |\boldsymbol{\beta}_j|^\gamma\}$.

Equivalently,

$\widehat{\beta}_{\mathbf{brdg}} = \arg\min_{\beta}\{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\}$ subject to $\sum_{j=1}^{p} |\boldsymbol{\beta}_j|^\gamma \leq \boldsymbol{t}$, with $\boldsymbol{t} \geq \mathbf{0}$.

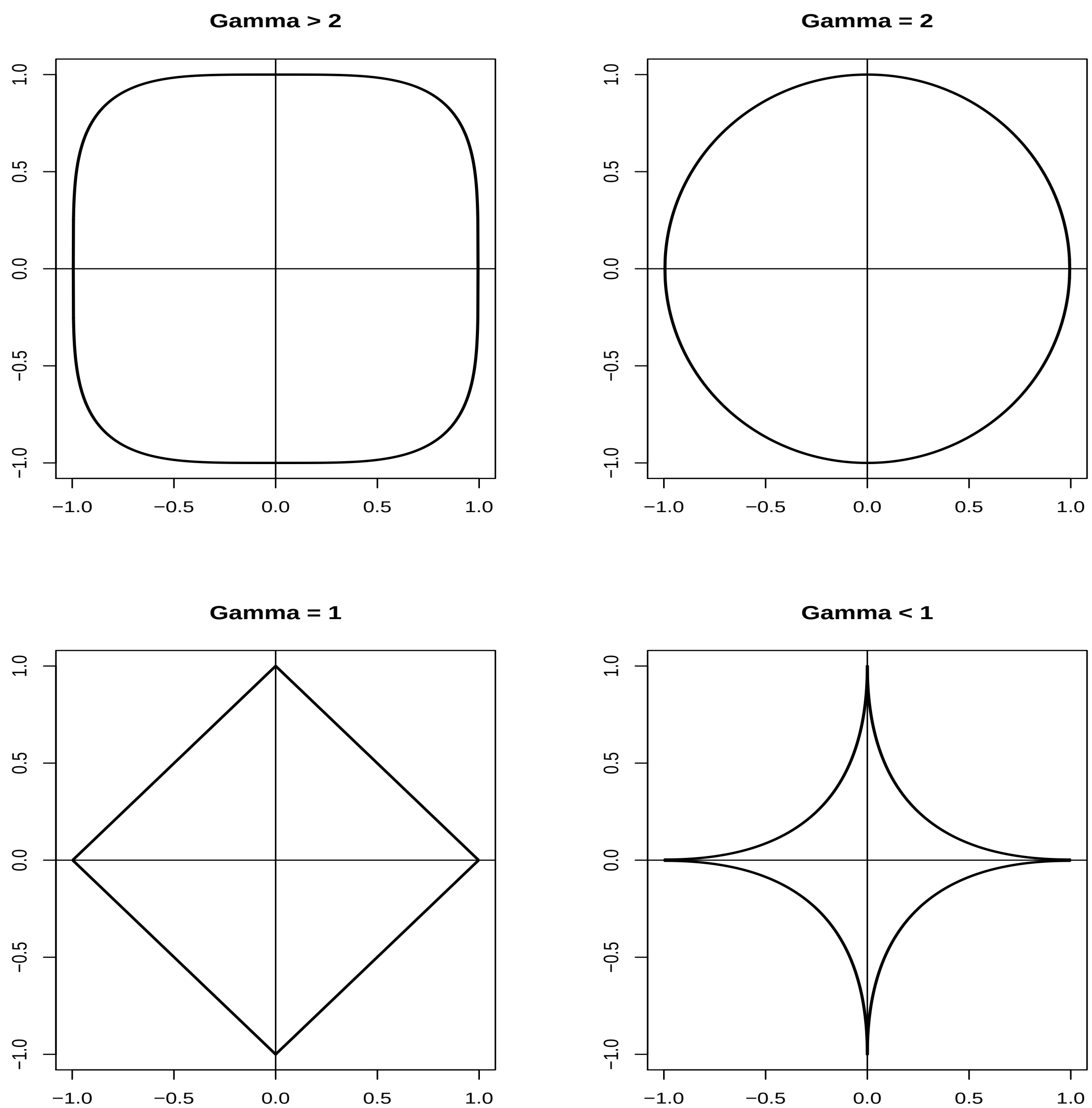
$$\mathbf{var}(\widehat{\beta}_{\mathbf{brdg}}) \leq \mathbf{var}(\widehat{\beta}_{\mathbf{ols}})$$

Bridge – generalization of ridge

- $\diamond\ \gamma = 2$, ridge;
- $\diamond\ \gamma = 1$, lasso (Tibshirani 1996).

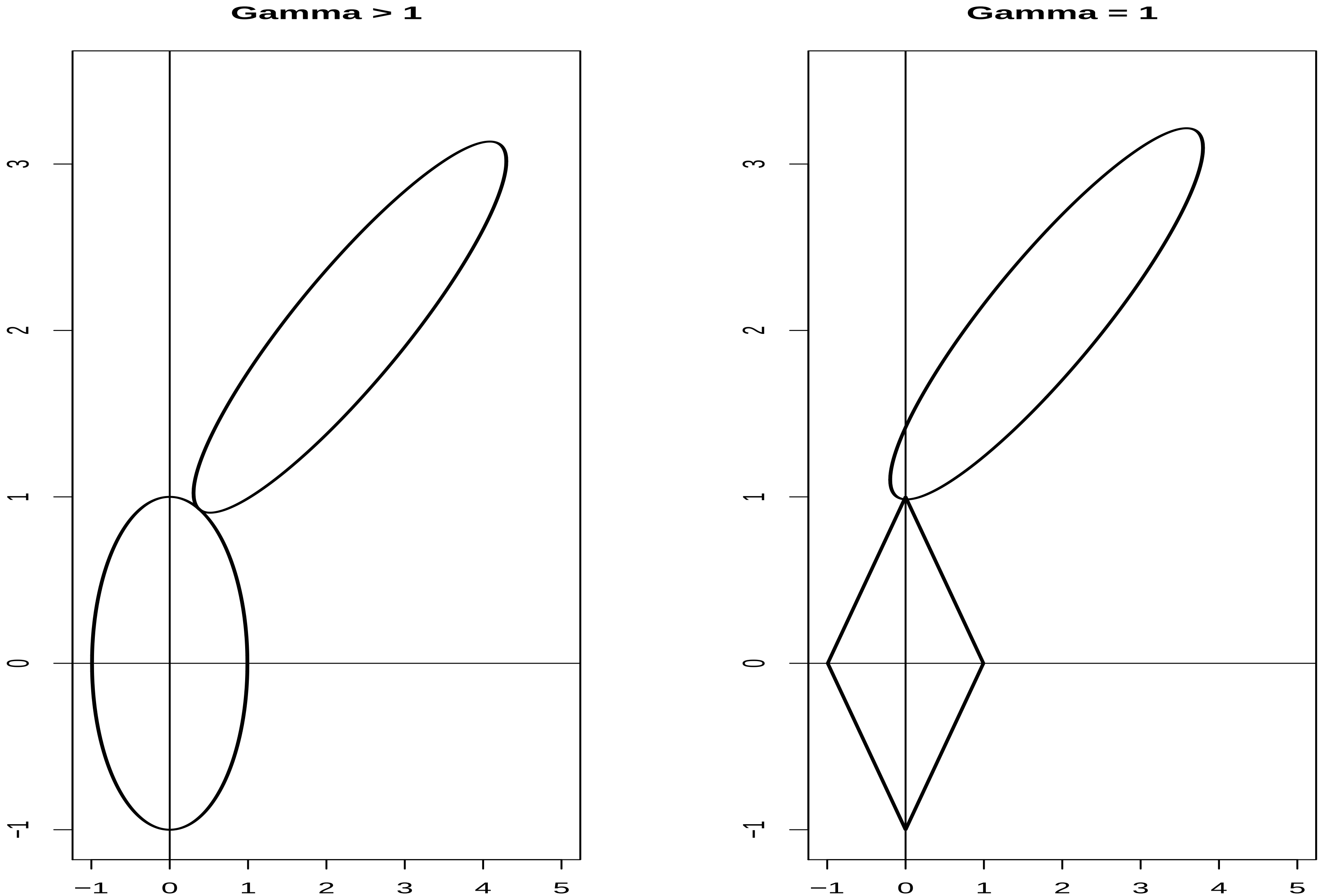
L^{γ} PENALTY

Constraint area for different values of $\gamma > 0$.



L^γ PENALTY

Variable selection property of lasso $\widehat{\beta}_j = 0$.



Computation for bridge $\gamma > 1$

- $\gamma = 2$: closed form.
- $\gamma > 1$: modified Newton-Raphson (Fu 1998), complex!

Notations:

$RSS = (y - X\beta)^T(y - X\beta)$, $S_j = \partial RSS/\partial \beta_j$,
$d(\beta_j, \lambda, \gamma) = \lambda\gamma|\beta_j|^{\gamma-1}\text{sign}(\beta_j)$, $l_j = S_j + d(\beta_j, \lambda, \gamma)$.
Solve system of equations:

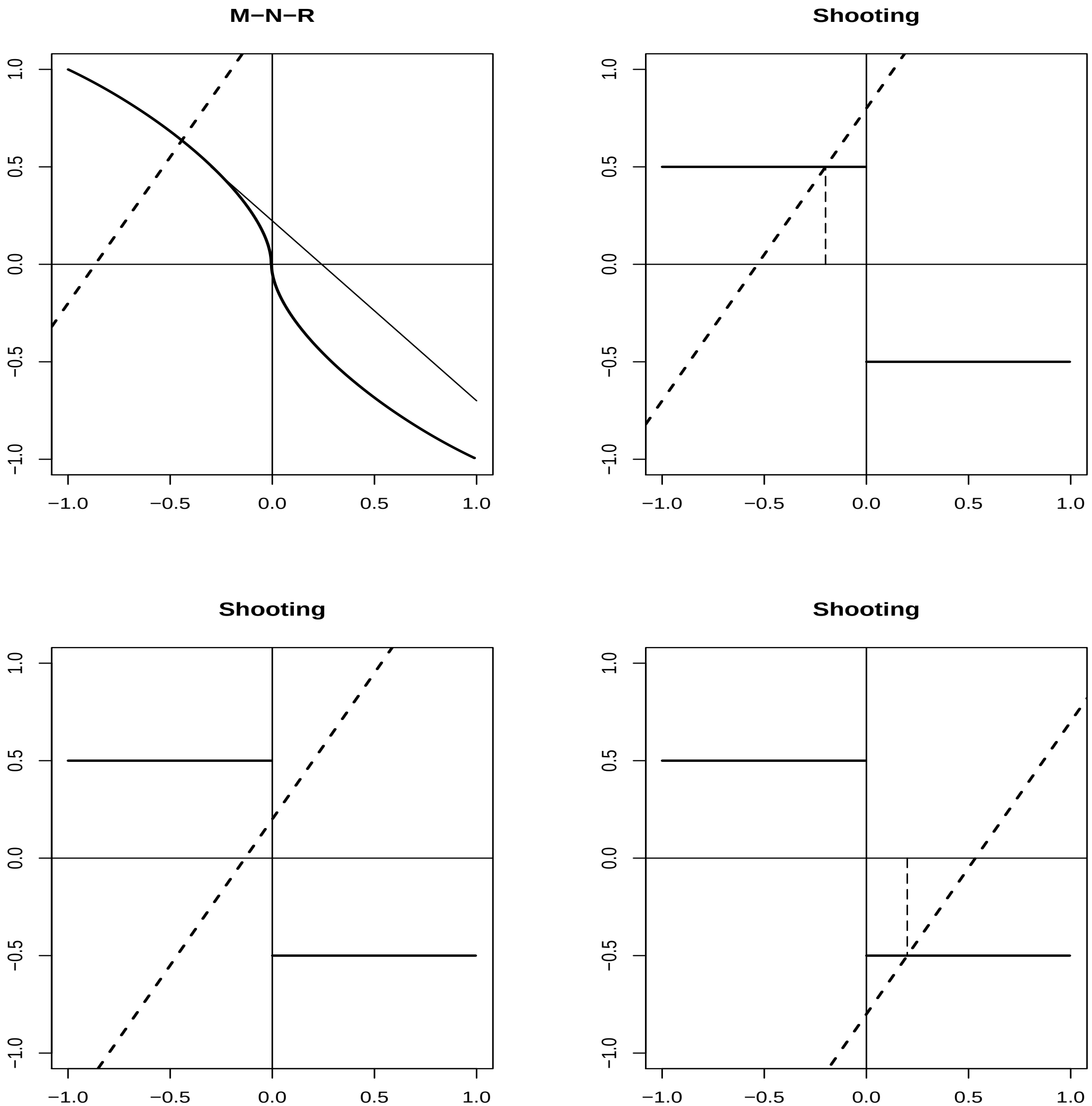
$$\begin{cases} l_1(\beta, X, y, \lambda, \gamma) = 0, \\ \quad \cdots \\ l_p(\beta, X, y, \lambda, \gamma) = 0. \end{cases} \tag{1}$$

No closed form. Use N–R. $\beta_j^{new} = \beta_j^{old} - [\partial l_j/\partial \beta_j]^{-1} l_j$

Modify N-R since convexity changes at $\beta_j = 0$ for $1 < \gamma < 2$.

L^γ PENALTY

M-N-R ($\gamma > 1$) and shooting algorithm ($\gamma = 1$).



Computation for lasso $\gamma = 1$

- Combined quadratic programming (Tibshirani 1996).
  Quadratic programming:
  $$\min(\boldsymbol{y} - \boldsymbol{X\beta})^T(\boldsymbol{y} - \boldsymbol{X\beta}) \text{ subject to } \boldsymbol{v}^T\boldsymbol{\beta} \geq \boldsymbol{0}.$$
  Constraint $\sum_{j=1}^{p} |\beta_j| \leq t$ is equivalent to
  $\sum_{j=1}^{p} w_j\beta_j \leq t$ with $w_j = \pm 1$.
  Total combinations of 2^p weights w_j. Complicated!

- Shooting algorithm (Fu 1998).
  Take limit $\gamma \to 1+$:
  not computationally – more complicated;
  but theoretically – iteration with simple closed form.

L^γ PENALTY

Theorem 1

If S_j is contin. diff., Jacobian $\partial S/\partial \beta$ pos-semi-def., then

1. $\widehat{\beta}(\lambda, \gamma)$ is unique and contin. in (λ, γ).
2. $\lim_{\gamma \to 1+} \widehat{\beta}(\lambda, \gamma)$ exists for fixed $\lambda > 0$.
3. $\lim_{\gamma \to 1+} \widehat{\beta}(\lambda, \gamma) = \widehat{\beta}(\lambda, 1)$, the lasso estimator for L–S.

Implication

1. Penalty (shrinkage) models do not need joint likelihood.
Only Jacobian $\partial S/\partial \beta$ condition (p.s.d.). Potential extension!
2. If joint likelihood exists, the extension works perfectly.

Shooting algorithm for lasso

1). Start with $\widehat{\beta}^{(0)} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)$.

2). At step m, for $j = 1, \ldots, p$, let $s_0 = S_j(0, \widehat{\beta}^{(-j)}, X, y)$ and x_j be the j–th column vector of X. Set

$$\widehat{\beta}_j = \begin{cases} \frac{\lambda - s_0}{2x_j^T x_j} & \text{if} \quad s_0 > \lambda \\ 0 & \text{if} \quad |s_0| \leq \lambda \\ \frac{-\lambda - s_0}{2x_j^T x_j} & \text{if} \quad s_0 < -\lambda \end{cases}$$

Form a new estimator $\widehat{\beta}^{(m)} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)$ after updating all $\widehat{\beta}_j$.

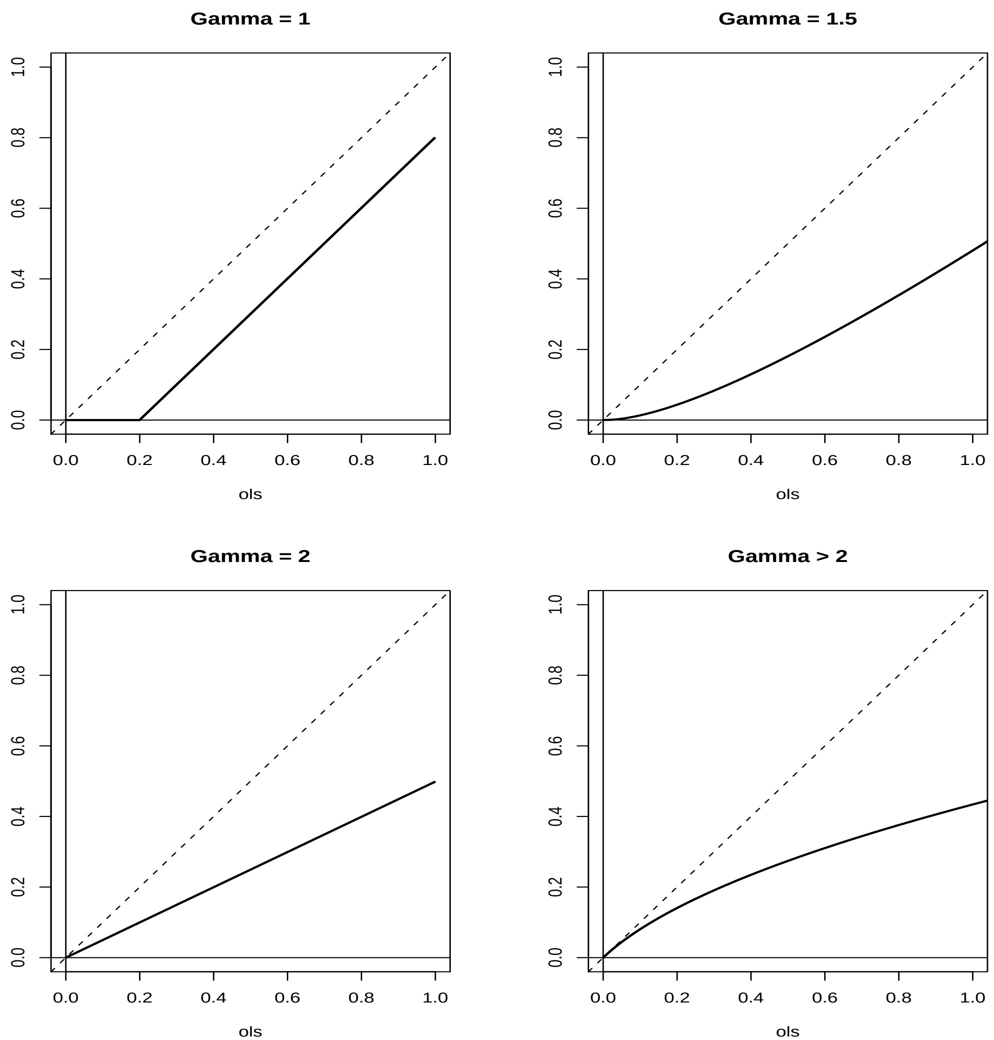
3). Repeat step 2) until convergence of $\widehat{\beta}^{(m)}$.

Convergence of algorithms

Let $\boldsymbol{G}(\boldsymbol{\beta}; \boldsymbol{\lambda}, \boldsymbol{\gamma}) = (\boldsymbol{y} - \boldsymbol{X\beta})^{\boldsymbol{T}}(\boldsymbol{y} - \boldsymbol{X\beta}) + \boldsymbol{\lambda}\sum_{\boldsymbol{j}} |\boldsymbol{\beta_j}|^\gamma$ for given $\lambda > 0$ and $\gamma \geq 1$. $\boldsymbol{G}(\boldsymbol{\beta}; \boldsymbol{\lambda}, \boldsymbol{\gamma})$ is convex and is minimized at finite $\boldsymbol{\beta} = \boldsymbol{\beta_0}$. Each step of updating $\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}}$ through either M-N-R algorithm or the shooting algorithm decreases the function $\boldsymbol{G}(\boldsymbol{\beta}; \boldsymbol{\lambda}, \boldsymbol{\gamma})$. Thus the estimator $\widehat{\boldsymbol{\beta}}_{\boldsymbol{m}}$ converges.

Orthonormal matrix X: $X^TX = I$

Coordinate: $\hat{\beta}_{\mathbf{brdg}} = \hat{\beta}_{\mathbf{ols}} - \lambda\gamma/2|\hat{\beta}_{\mathbf{brdg}}|^{\gamma-1}\mathbf{sign}(\hat{\beta}_{\mathbf{brdg}})$



Variance of bridge estimator

• $\gamma > 1$, complex closed form: no zero-valued coordinates. (Fu 1998).

$$\mathbf{var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{D}(\widehat{\boldsymbol{\beta}})|_{\boldsymbol{y_0}})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{D}(\widehat{\boldsymbol{\beta}})|_{\boldsymbol{y_0}})^{-1}\boldsymbol{\sigma}^{\mathbf{2}}, \quad (2)$$

$\boldsymbol{D}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\lambda}\gamma(\gamma - 1)/2\,\mathbf{diag}(|\boldsymbol{\beta_j}|^{\gamma-\mathbf{2}})$.

$\boldsymbol{y_0}$ is some point in the sample space.

Difficult to use for lasso due to $\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}} = \mathbf{0}$.

Variance of bridge estimator

• $\gamma = 1$, difficulty: zero-valued coordinates.

Method in Tibshirani (1996):

$$\mathrm{var}(\widehat{\beta}) = (X^T X + \lambda W^-)^{-1} X^T X (X^T X + \lambda W^-)^{-1} \sigma^2, \qquad (3)$$

where $W = \mathrm{diag}(|\beta_1|, \ldots, |\beta_p|)$

Method in Osborne (2000);

Let

$$W = \frac{X^T(y - X\widehat{\beta})(y - X\widehat{\beta})^T X}{\|\widehat{\beta}\|_1 \|X^T(y - X\widehat{\beta})\|_\infty}$$

$$\mathrm{var}(\widehat{\beta}) = (X^T X + W)^{-1} X^T X (X^T X + W)^{-1} \sigma^2. \qquad (4)$$

Comparison between two methods

• (3) is zero for $\widehat{\beta}_j = 0$, while (4) is non-zero.

• However, if set tuning parameter $\lambda > 0$ large, all $\beta_j = 0$. Then no variability. Hence variance should be zero.
(3) is acceptable, but (4) still non-zero.

• All the above methods are approximations. No exact results except for $\gamma = 2$.

• Bootstrap method usually yields good estimation.

Table 1. Analysis of prostate cancer data.

Predictor	$\widehat{\beta}^a$	SE by (1)	SE by (2)	$\widehat{\beta}^b$	SE by bootstrap
Intercept	2.478	0.072	0.072	2.478	0.072
lcavol	0.559	0.079	0.101	0.618	0.103
lweight	0.097	0.060	0.081	0.190	0.076
age	0	0	0.079	-0.048	0.046
lbph	0	0	0.080	0.103	0.066
svi	0.156	0.071	0.097	0.245	0.087
lcp	0	0	0.125	0	0.068
gleason	0	0	0.114	0	0.047
pgg45	0	0	0.123	0.063	0.056

[a] Osborne (2000) ($t = 0.8114$). [b] Fu (1998) ($\lambda = 7.2$).

Shrinkage trace

— Parameter estimates change with a special tuning parameter: standard shrinkage rate, $0 \leq s \leq 1$.

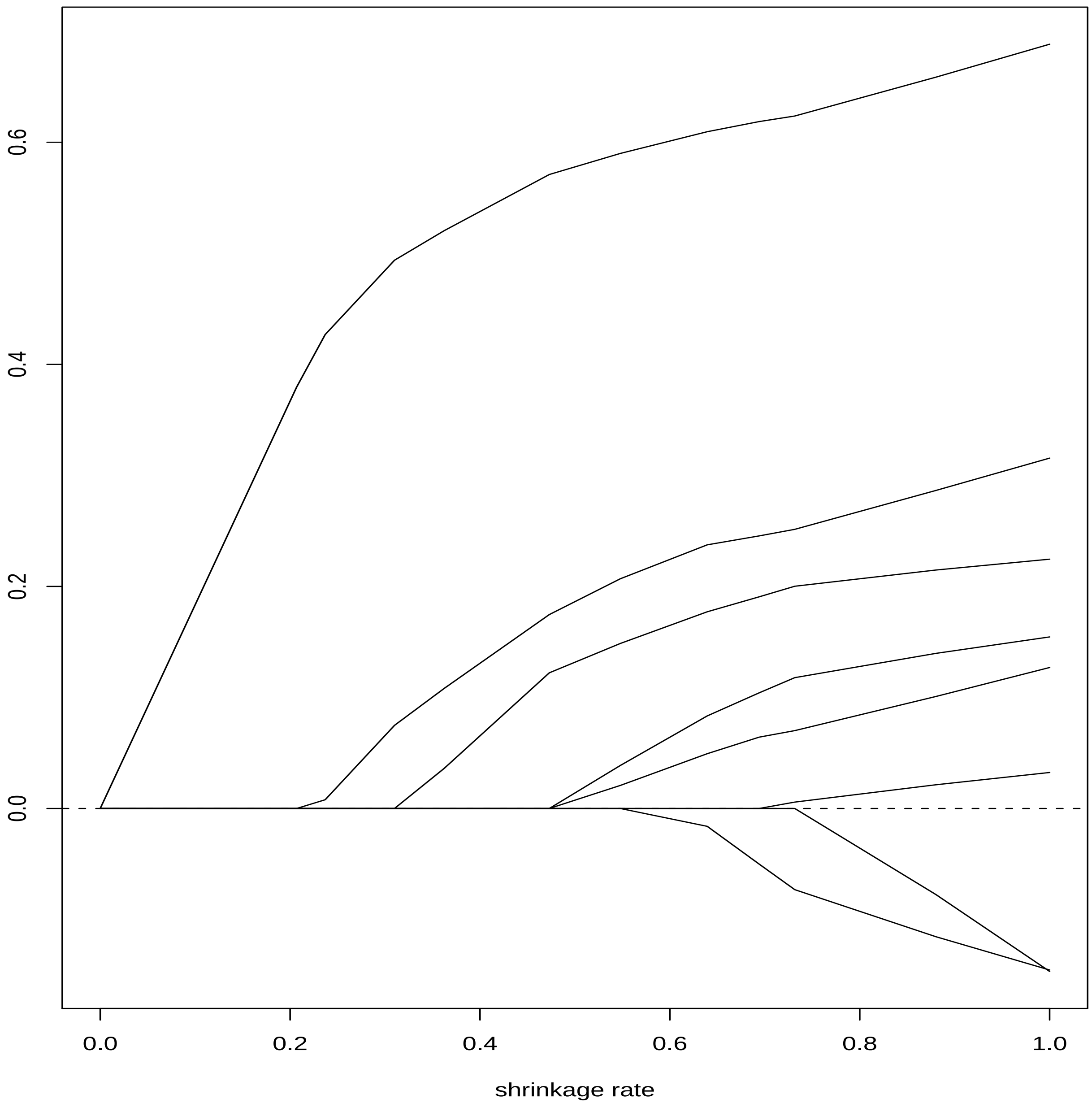
$$s = \frac{\|\beta_j(\lambda, \gamma)\|_\gamma}{\|\beta_j(\lambda = 0, \gamma)\|_\gamma},$$

where $\|\cdot\|_\gamma$ is the L^γ norm of a p-vector.

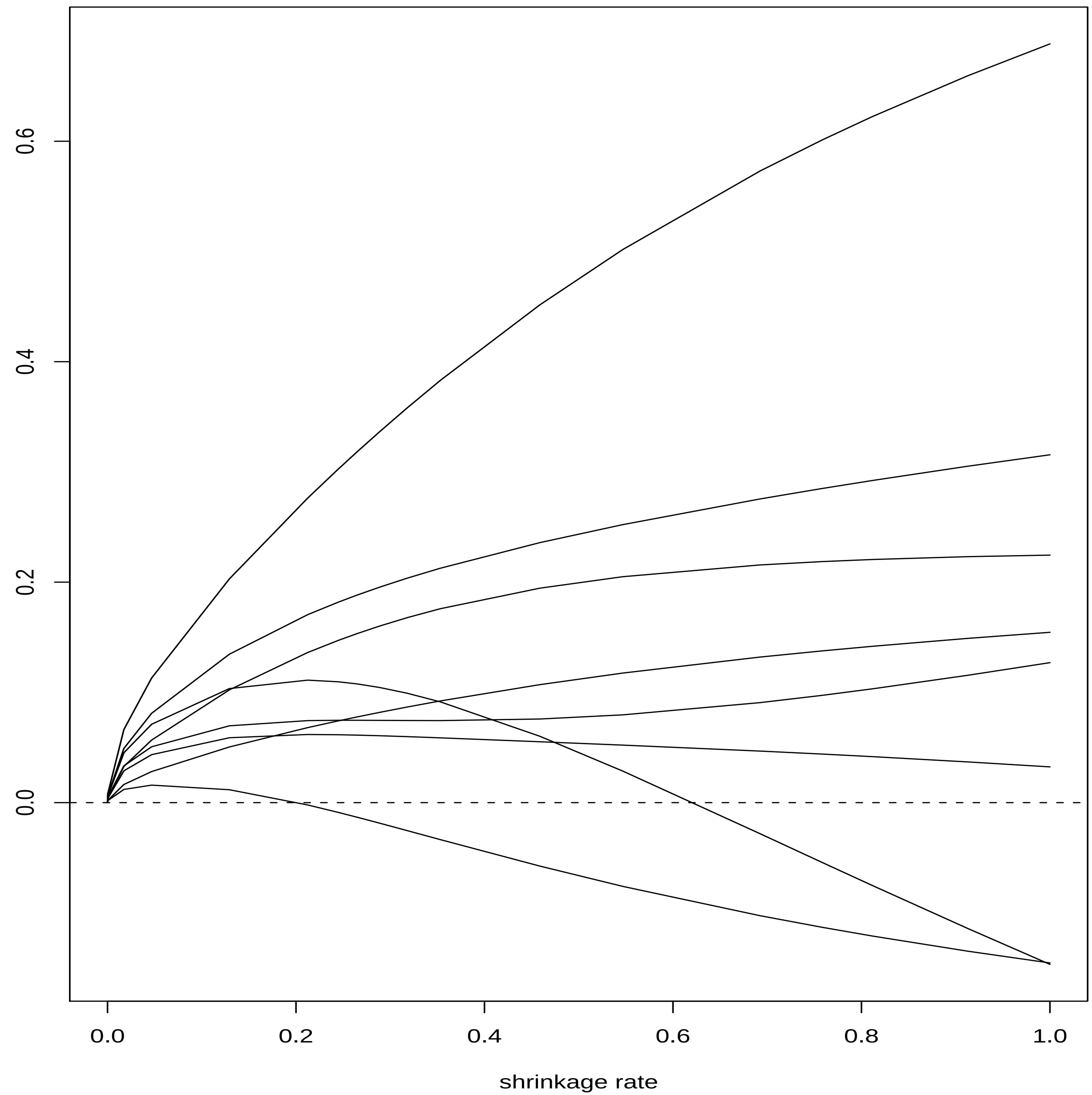
$s = 0$, full shrinkage.

$s = 1$, no shrinkage.

L^γ Penalty – Lasso Shrinkage Trace



L^γ Penalty – Ridge Shrinkage Trace



Seemingly contradictory results:

Given data $(\boldsymbol{X}, \boldsymbol{y})$: $\quad \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{\sigma}^{\mathbf{2}})$.

Let $(\boldsymbol{x}, \boldsymbol{y})$ be an arbitrary point in sample space.

◇ Prediction error increases with collinearity.

$$\begin{aligned} \mathbf{PSE}(\boldsymbol{x}) &= \boldsymbol{\sigma}^{\mathbf{2}} + \mathbf{MSE}(\boldsymbol{x}) = \boldsymbol{\sigma}^{\mathbf{2}}[\mathbf{1} + \boldsymbol{x}^{\boldsymbol{T}}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{x}] \\ &= \boldsymbol{\sigma}^{\mathbf{2}} + \boldsymbol{x}^{\boldsymbol{T}}\mathbf{var}(\widehat{\boldsymbol{\beta}})\boldsymbol{x}. \end{aligned}$$

◇ Prediction error at given data points is constant.

$$\begin{aligned} &\frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\mathbf{1}}^{\boldsymbol{n}}\mathbf{PSE}(\boldsymbol{x}_{\boldsymbol{i}}) = \boldsymbol{\sigma}^{\mathbf{2}}[\mathbf{1} + \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\mathbf{1}}^{\boldsymbol{n}}\boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{T}}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{x}_{\boldsymbol{i}}] \\ &= \boldsymbol{\sigma}^{\mathbf{2}}[\mathbf{1} + \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\mathbf{1}}^{\boldsymbol{n}}\mathbf{tr}\{(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{x}_{\boldsymbol{i}}\boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{T}}\}] = \boldsymbol{\sigma}^{\mathbf{2}}(\mathbf{1} + \frac{\boldsymbol{p}}{\boldsymbol{n}}), \end{aligned}$$

where $\boldsymbol{x}_{\boldsymbol{i}}$ are row vectors of matrix $\boldsymbol{X}$.

◇ In fact, $\mathbf{E}[\mathbf{PSE}(\boldsymbol{x})] = \boldsymbol{\sigma}^{\mathbf{2}}(\mathbf{1} + \frac{\boldsymbol{p}}{\boldsymbol{n}})$.

Collinearity increases variability of PSE:

Proposition (Fu 2005)

Assume existence of two moments $\mathbf{E}(\boldsymbol{x}\boldsymbol{x}^T)$ and $\mathbf{E}(\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{x}\boldsymbol{x}^T)$. The expectation $\mathbf{E}\{\mathbf{PSE}(\boldsymbol{x})\}$is independent of the collinearity for large samples with $\mathbf{E}\{\mathbf{PSE}(\boldsymbol{x})\} \sim \boldsymbol{\sigma}^2(1+\boldsymbol{p}/\boldsymbol{n})$. The variance $\mathbf{var}\{\mathbf{PSE}(\boldsymbol{x})\}$ increases with the collinearity as the smallest eigenvalue of matrix $\boldsymbol{X}^T\boldsymbol{X}$ decreases to 0.

$$\mathbf{var}\{\mathbf{PSE}(\boldsymbol{x})\} = \mathbf{E}[\{\mathbf{PSE}(\boldsymbol{x})\}^2] - [\mathbf{E}\{\mathbf{PSE}(\boldsymbol{x})\}]^2$$

$$\sim \frac{\boldsymbol{\sigma}^4}{\boldsymbol{n}^2}[\mathbf{tr}\{\boldsymbol{V}^{-1}\mathbf{E}(\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{V}^{-1}\boldsymbol{x}\boldsymbol{x}^T)\} - \boldsymbol{p}^2],$$

where $\boldsymbol{V} = \boldsymbol{X}^T\boldsymbol{X}/\boldsymbol{n}$.

Collinearity increases variability of PSE:

Assume $\boldsymbol{V} = \mathrm{diag}(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_p)$ with $\boldsymbol{\lambda}_1 \geq \ldots \geq \boldsymbol{\lambda}_p > 0$, without loss of generality.

Let $\boldsymbol{U} = \mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{V}^{-1}\boldsymbol{x}\boldsymbol{x}^T) - \mathrm{E}(\boldsymbol{\lambda}_1^{-1}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{x}\boldsymbol{x}^T)$, psd.

$\boldsymbol{U}$ and $\boldsymbol{V}$ can be diagonalized simultaneously. $\boldsymbol{V}^{-1}\boldsymbol{U}$ is psd.

$$\begin{aligned}
\mathrm{tr}\{\boldsymbol{V}^{-1}\mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{V}^{-1}\boldsymbol{x}\boldsymbol{x}^T)\} &\geq \mathrm{tr}\{\boldsymbol{V}^{-1}\boldsymbol{\lambda}_1^{-1}\mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{x}\boldsymbol{x}^T)\} \\
&= \boldsymbol{\lambda}_1^{-1}[\boldsymbol{\lambda}_1^{-1}\boldsymbol{c}_1 + \cdots + \boldsymbol{\lambda}_p^{-1}\boldsymbol{c}_p] \\
&> \boldsymbol{\lambda}_1^{-1}\boldsymbol{\lambda}_p^{-1}\boldsymbol{c}_p \longrightarrow \infty \text{ as } \boldsymbol{\lambda}_p \longrightarrow 0,
\end{aligned}$$

where $\boldsymbol{c}_1, \cdots, \boldsymbol{c}_p > 0$ are the elements on the main diagonal of matrix $\mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{x}\boldsymbol{x}^T)$ and are independent of matrix $\boldsymbol{V}$.

References

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

Fu, W.J. (1998). Penalized regressions: the Bridge versus the Lasso, *J. Comp. Grap. Statist.* 7: 397-416.

Fu, W.J. (2005). Prediction error with collinearity, *Comm. Statist. - Theor. Meth.*, in press.

Gruber, M.H.J. (1990). *Regression Estimators: A Comparative Study*, Academic Press, Boston.

Hoerl, A.E. and Kennard, R.W. (1970a). Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, 12:55-67.

Hoerl, A.E. and Kennard, R.W. (1970b). Ridge regression: applications to nonorthogonal problems, *Technometrics*, 12:69-82.

James, W. and Stein, C (1961). Estimation with quadratic loss, *Proc. Fourth Berkeley Symp. Math. Statist Prob.* 1, 311-319.

Land, K.C. McCall, P.L. and Cohen, L.E. (1990). Structural covariates of homicide rates: are there any invariances across time and social space? *American Journal of Sociology*, 95, 922-963.

Miller, A.J. (1990). *Subset Selection in Regression,* Chapman and Hall, New York.

Osborne, M. R., Presnell, B. and Turlach, B. A. (2000). On the LASSO and its dual. *Journal of Computational and Graphical Statistics* **9**, 319-337.

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

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