Absolute Penalty and B-spline-based Shrinkage Estimation in Partially Linear Models

Enayetur Raheem

University of Windsor / Windsor-Essex County Health Unit
Windsor, Ontario, Canada

International Workshop on High Dimensional Data Analysis
Fields Institute, June 9-11, 2011

With
K. A. Doksum (Wisconsin-Madison) and S Ejaz Ahmed (Windsor)

Outline

- Overview of shrinkage estimation
- Motivational Examples
- Asymptotic properties of shrinkage estimators
- Monte Carlo study
- Conclusion
- Software for shrinkage estimation

Brief introduction to shrinkage estimation

- Dates back to Stein (1956); then improved by James and Stein (1961).
- Supose $\boldsymbol{\theta}_{p\times 1}$ is an unknown parameter vector, and $\mathbf{y}$ be a vector of observations of $\boldsymbol{\theta}$, s.t.

$$\mathbf{y} \sim N(\boldsymbol{\theta}, \sigma^2 I)$$

  We want an estimate $\hat{\boldsymbol{\theta}}$ based on $\mathbf{y}$
- Since the noise has zero mean, we get using the concept of Least-squares

$$\hat{\boldsymbol{\theta}_{LS}} = \bar{\mathbf{y}}$$

- Stein (1956) demonstrated that, in terms of a quardadic loss function, $E(||\theta - \hat{\theta}||^2)$, this approach is suboptimal. The result became known as Stein's phenomenon.

James-Stein Estimator

- If σ^2 is known, James-Stein showed that

$$\hat{\boldsymbol{\theta}}_{JS} = \left\{ 1 - \frac{(p-2)\sigma^2}{||\bar{\mathbf{y}}||^2} \right\} \bar{\mathbf{y}}$$

  domainates $\hat{\boldsymbol{\theta}}_{LS}$ for $p \geq 3$

  $\Rightarrow$ JS estimator always achieves lower MSE than the LS estimator.

Shrinking towards ν

$$\hat{\boldsymbol{\theta}}_{JS} = \left\{1 - \frac{(p-2)\sigma^2}{||\bar{\mathbf{y}}||^2}\right\}\bar{\mathbf{y}}$$

- Notice that if $(p-2)\sigma^2 < ||\bar{\mathbf{y}}||^2$, this estimator shrinks the natural estimator $\bar{\mathbf{y}}$ towards zero.
- For any fixed vector $\boldsymbol{\nu}$ of length p, there exists a JS estimator that shrinks towards $\boldsymbol{\nu}$, which is

$$\hat{\boldsymbol{\theta}} = \left\{1 - \frac{(p-2)\sigma^2}{||(\bar{\mathbf{y}} - \boldsymbol{\nu})||^2}\right\}(\bar{\mathbf{y}} - \boldsymbol{\nu}) + \boldsymbol{\nu}$$

  for any $\boldsymbol{\nu}$

Optimal choice of $\boldsymbol{\nu}$

- A natural question is whether the improvement over the usual estimator is independent of the choice of $\boldsymbol{\nu}$
- Improvement is large if $||\boldsymbol{\theta} - \boldsymbol{\nu}||$ is small

Usefulness

- Shrinkage estimators are more efficient than the classical estimators for $p \geq 3$ (in terms of a quadratic risk function, e.g., MSE)
- Ensures improved prediction accuracy

Motivating Examples

[Mroz (Econometrica, 1987)] used a sample of 1975 PSID[1] labour supply data to systematically study several theoretic and statistical assumptions used in many empirical models of female labour supply.
This data set was used in numerous Econometric studies and books.

- [Wooldridge, 2003] used this data in his book to demonstrate many applications of regression models.
- [Long, 1997] fitted parametric logistic regresssion model to this data.
- [Fox, 2002] used this data for a semiparametric logistic regression.

[1] Panel Study on Income Dynamics (PSID), University of Michigan. The data is in public domain and freely availale from `http://ideas.repec.org/s/boc/bocins.html`

Motivating Examples ...

- [Fox, 2005] commented that semiparamertic model may be used whererver there is reason to believe that one or more covariates enter the regression linearly. This could be known from prior studies, or there are prior reasons to believe so (althogh rare), or examination of the data might suggest a linear relationship for some covariates. A more general scenario is when some of the covariates are categorical and they enter in the model as dummy variables.
- [Engle (JASA, 1986)] considered such a PLM where demand for electricy was the outcome variable and four regions were entered as dummy variable while temparature was modelled nonparametrically.

Objective

- Motivated by [Engle (JASA, 1986)] and [Fox, 2002], we fit a semiparametric regression model with a continuous outcome variable and demonstrate shrinkage estimation using PSID data.
- We first show that semiparametric modeling is appropriate in our case.

Partially Linear Model

- A partially linear regression model:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + g(t_i) + \varepsilon_i, i = 1, \ldots, n, \tag{1}$$

where y_i's are responses,
$\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ and $t_i \in [0, 1]$ are design points,
$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is an unknown parameter vector,
$g(\cdot)$ is an unknown real-valued function defined on $[0, 1]$, and
ε_i's are unobservable random errors.

Related earlier work

- Extension of Ahmed et al. (2007)

Related earlier work

- Extension of Ahmed et al. (2007)

Objective

- We consider estimating $\boldsymbol{\beta}$ based on $g(\cdot)$ approximated by a B-spline series.

Data and Variables

- The female labour supply data consist of 753 observations on 19 variables. Data were collected from married white women between the ages 30 and 60 in 1975. Of them, 428 were working at some time during the year 1975.
- Similar to [Mroz (Econometrica, 1987)], we consider wife's annual hours of work (`hours`) as the response variable.
- We only used the portion of the data when the women were in labour force. Thus, we had 428 cases (rows) in our working data.

Table: Description of Variables in the Model for Working Women.

Covariates	Description	Remarks
`hours`	Hrs worked in 1975	Min=12, max=4950, med=1303
`age`	Age of woman	Min=30, max=60, med=42
`nwifeinc`	Non-wife income	Income in thousands
`k5`	Kids five and under	0-1, a few 2's and 3's, factor
`k618`	Kids six to 18	0-4, few >4, factor variable
`wc`	W attended college?	1 (if `educ` $>$ 12), else 0
`hc`	H attended college?	1 (if `huseduc` $>$ 12), else 0
`unem`	Unemployment rate	Min=3, max=14, median=7.5
`mtr`	Marginal tax rate	Min=0.44, max=0.94, med=0.69
`exper`	Labour market exp	Min=0, max=38, median=12

Algorithm

- When we have prior information about certain covariates: shrinkage estimators are directly obtain by combining the full and sub-model estimates.

if *a priori* information is not available, it take a two step approach

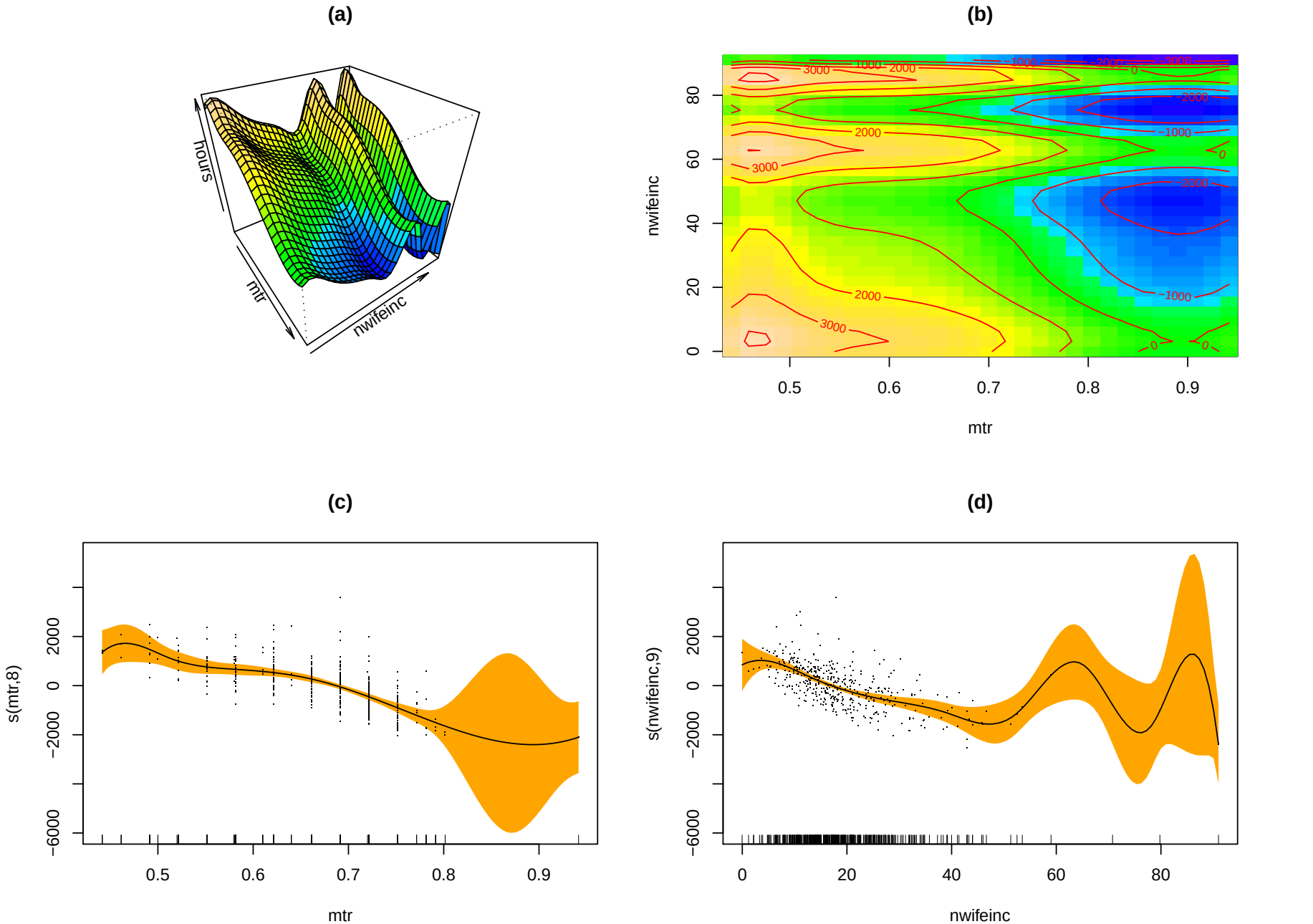
Step 1 a set of covariates are selected based on model selection criterion (AIC, BIC etc.) that form the main covariate-set. Consequently, the remaining covariates are taken as nuisance subset–forming the restriction on the full model.

Step 2 Full- and the sub-model estimates are combined in a way that minimizes the quadratic risk function.

Table: Selection of covariates by AIC, BIC and lasso.

Models	Selected Variables
Full Model	`wc`, `nwifeinc`, `mtr`, `exper`, `unem`, `k5`, `age`, `k618`, `hc`
AIC	`wc`, `nwifeinc`, `mtr`, `exper`, `unem`, `k5`, `age`
BIC	`wc`, `nwifeinc`, `mtr`, `exper`
Lasso	`wc`, `nwifeinc`, `mtr`, `exper`, `unem`, `k5`, `age`, `hc`

- Our analyses show that `nwifeinc` has significant nonlinear relationship ($p = 0.0065$) with hours of work.


(a)
hours
mtr
nwifeinc
(b)
nwifeinc
mtr
3000
2000
1000
0
−1000
−2000
(c)
s(mtr,8)
mtr
(d)
s(nwifeinc,9)
nwifeinc

Full- and Sub-model

Finally, with the inclusion of a nonparametric part, our caldidate full- and sub-models are listed below.

Full-Model:

$$\begin{aligned}\texttt{hours} = \texttt{wc} + g(\texttt{nwifeinc}) + \texttt{mtr} + \ \texttt{exper} + \texttt{unem} \\ + \texttt{k5} + \texttt{age} + \texttt{k618} + \texttt{hc}\end{aligned}$$

Sub-Model:

$$\texttt{hours} = \texttt{wc} + g(\texttt{nwifeinc}) + \texttt{mtr} + \ \texttt{exper}$$

Here $g(\cdot)$ denotes a covariate estimated by B-spline basis function.

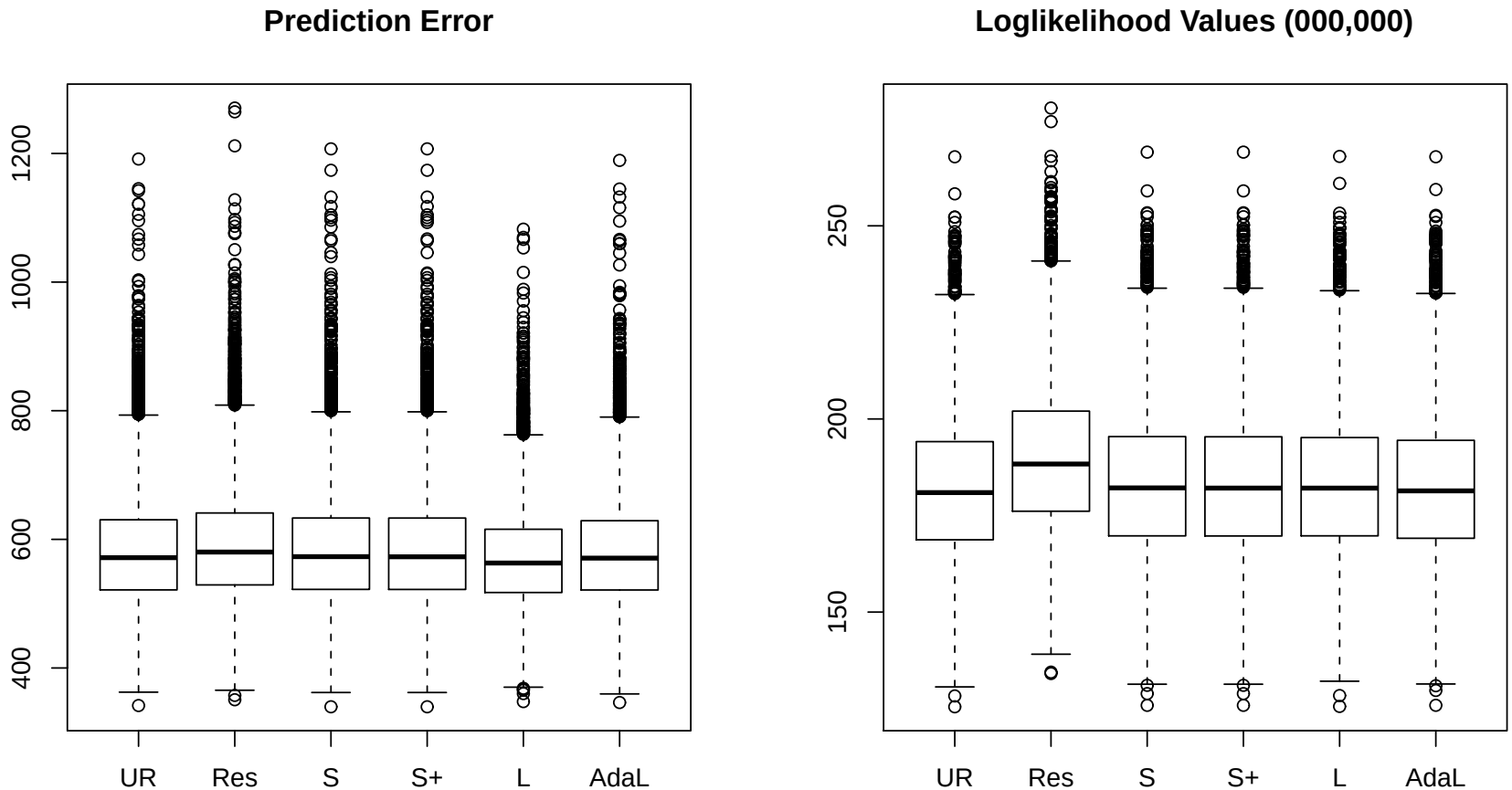


Figure: Comparison of the estimators through prediction errors and loglikelihood values.

THEORETICAL RESULTS:Parameter Estimation

Estimation of $g(\cdot)$ by B-spline

- Let k be an integer larger than or equal to ν where ν will be defined in the Assumption (2)
- Further, let $S_{m_n,k}$ be the class of functions $s(\cdot)$ on $[0,1]$ with the following properties:
  - (i) $s(\cdot)$ is a polynomial of degree k on each of the sub-intervals $\left[\frac{(i-1)}{m_n}, \frac{i}{m_n}\right], i = 1, \ldots, m_n$, where m_n is a positive integer which depends on n.
  - (ii) $s(\cdot)$ is $(k-1)$ times differentiable.
- Then, $S_{m_n,k}$ is called the class of all splines of degree k with m_n-equispaced knots.

Parameter Estimation...

Consequently,

- $S_{m_n,k}$ has a basis of $m_n + k$ normalized B-spline $\{B_{m_nj}(\cdot) : j = 1, \ldots, m_n + k\}$, and
- $g(\cdot)$ can be approximated by a linear combination $\boldsymbol{\theta}'\mathbf{B}_{m_n}(\cdot)$ of the basis, where $\boldsymbol{\theta} \in \mathcal{R}^{m_n+k}$ and $\mathbf{B}_{m_n}(\cdot) = (B_{m_n1}(\cdot), \ldots, B_{m_n,m_n+k}(\cdot))'$. See de Boor (2001)

- Now, replacing $g(\cdot)$ by $\mathbf{B}_{m_n}(\cdot)\boldsymbol{\theta}$, in model (1) we get

$$\boxed{y = \mathbf{x}\boldsymbol{\beta} + \mathbf{B}_{m_n}(t)\boldsymbol{\theta} + \varepsilon} \tag{2}$$

Full Model Estimation

- $(\hat{\boldsymbol{\beta}}, \hat{\theta})$ is obtained by minimizing

$$S_n(\boldsymbol{\beta}, \boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} \left[y_i - \mathbf{x}_i' \boldsymbol{\beta} - \boldsymbol{\theta}' \mathbf{B}_{m_n}(t_i) \right]^2, \qquad (3)$$

which gives

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{M}_{\mathbf{B}_{m_n}} \mathbf{X})^{-1} \mathbf{X}' \mathbf{M}_{B_{m_n}} \mathbf{Y}$$

and

$$\hat{\boldsymbol{\theta}} = (\mathbf{B}_{m_n}' \mathbf{B}_{m_n})^{-1} \mathbf{B}_{m_n}' (\mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\beta}}),$$

where, $\mathbf{Y} = (y_1, \ldots, y_n)'$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$,
$\mathbf{x}_s = (x_{1s}, \ldots, x_{ns})'$, $s = 1, \ldots, p$,
$\mathbf{M}_{\mathbf{B}_{m_n}} = \mathbf{I} - \mathbf{B}_{m_n} (\mathbf{B}_{m_n}' \mathbf{B}_{m_n})^{-1} \mathbf{B}_{m_n}'$ and
$\mathbf{B}_{m_n} = (B_{m_n}(t_1), \ldots, B_{m_n}(t_n))'$.

- The estimator $\hat{\boldsymbol{\beta}}$ is called a semiparametric least squares estimator (SLSE) of $\boldsymbol{\beta}$.

Uncertain Prior Information

- $\boldsymbol{\beta}$ in the linear part can be partitioned as $(\boldsymbol{\beta_1}, \boldsymbol{\beta_2})$
- $\boldsymbol{\beta_1}$ is the coefficient vector for main effects (e.g., treatment effect, genetic effects) and $\boldsymbol{\beta_2}$ is a vector for "nuisance" effects (e.g., age, laboratory).

Sub Model Estimation

Semi-Parametric Unrestricted Estimator (SURE)

$$\hat{\beta}_1^{UR} = (\mathbf{X}_1' \mathbf{M}_{\mathbf{B}_{mn}} \mathbf{M}_{\mathbf{B}_{mn}\mathbf{x}_2} \mathbf{M}_{\mathbf{B}_{mn}} \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{M}_{\mathbf{B}_{mn}} \mathbf{M}_{\mathbf{B}_{mn}\mathbf{x}_2} \mathbf{M}_{\mathbf{B}_{mn}} \mathbf{Y},$$

where
$\mathbf{X}_1$ is composed of the first p_1 column vectors of $\mathbf{X}$,
$\mathbf{X}_2$ is composed of the last p_2 column vectors of $\mathbf{X}$, and
$\mathbf{M}_{\mathbf{B}_{mn}\mathbf{x}_2} = \mathbf{I} - \mathbf{B}_{mn}\mathbf{X}_2(\mathbf{X}_2'\mathbf{B}_{mn}'\mathbf{B}_{mn}\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{B}_{mn}'$.

Sub Model Estimation

Semi-Parametric Unrestricted Estimator (SURE)

$$\hat{\beta}_1^{UR} = (\mathbf{X}_1' \mathbf{M}_{\mathbf{B}_{m_n}} \mathbf{M}_{\mathbf{B}_{m_n}\mathbf{x}_2} \mathbf{M}_{\mathbf{B}_{m_n}} \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{M}_{\mathbf{B}_{m_n}} \mathbf{M}_{\mathbf{B}_{m_n}\mathbf{x}_2} \mathbf{M}_{\mathbf{B}_{m_n}} \mathbf{Y},$$

where
$\mathbf{X}_1$ is composed of the first p_1 column vectors of $\mathbf{X}$,
$\mathbf{X}_2$ is composed of the last p_2 column vectors of $\mathbf{X}$, and
$\mathbf{M}_{\mathbf{B}_{m_n}\mathbf{x}_2} = \mathbf{I} - \mathbf{B}_{m_n}\mathbf{X}_2(\mathbf{X}_2'\mathbf{B}_{m_n}'\mathbf{B}_{m_n}\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{B}_{m_n}'$.

Semi-Parametric Restricted Estimator (SRE)

$$\hat{\beta}_1^{R} = (\mathbf{X}_1' \mathbf{M}_{\mathbf{B}_{m_n}} \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{M}_{\mathbf{B}_{m_n}} \mathbf{Y}.$$

Semiparametric Stein-Type Estimation

A semiparametric Stein-type estimator (SSTE) $\hat{\boldsymbol{\beta}}_1^S$ of $\boldsymbol{\beta}_1$ can be defined as

$$\hat{\boldsymbol{\beta}}_1^{\mathrm{S}} = \hat{\boldsymbol{\beta}}_1^{\mathrm{R}} + (\hat{\boldsymbol{\beta}}_1^{\mathrm{UR}} - \hat{\boldsymbol{\beta}}_1^{\mathrm{R}}) \left\{1 - (p_2 - 2)\psi_n^{-1}\right\}, \quad p_2 \geq 3.$$

where

$$\psi_n = \frac{n}{\hat{\sigma}_n^2} \hat{\boldsymbol{\beta}}_2' \mathbf{X}_2' \mathbf{B}_{m_n}' \mathbf{M}_{\mathbf{B}_{m_n}\mathbf{x}_2} \mathbf{B}_{m_n} \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2,$$

with

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}} - \mathbf{B}_{m_n}'(t_i) \hat{\boldsymbol{\theta}})^2.$$

Positive-part Semiparametric Stein-Type Estimator

A PSSTE has the form

$$\hat{\boldsymbol{\beta}}_1^{\mathrm{S}+} = \hat{\boldsymbol{\beta}}_1^{\mathrm{R}} + (\hat{\boldsymbol{\beta}}_1^{\mathrm{UR}} - \hat{\boldsymbol{\beta}}_1^{\mathrm{R}}) \left\{1 - (p_2 - 2)\psi_n^{-1}\right\}^+, \quad p_2 \geq 3.$$

where $z = \max(0, z)$.

Absolute Penalty Estimator

The lasso estimates are obtained as

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\text{argmin}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

Tibshirani (1996).

Asymptotics

Assumption 1

There exist bounded functions $h_s(\cdot)$ over $[0,1]$, $s = 1, \ldots, p$, such that

$$x_{is} = h_s(t_i) + u_{is}, \;\; i = 1, \ldots, n, s = 1, \ldots, p, \qquad (a)$$

where $u_i = (u_{i1}, \ldots, u_{ip})'$ are real vectors satisfying

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} u_{ik} u_{ij}}{n} = b_{kj}, \quad \text{for } k = 1, \ldots, p, \; j = 1, \ldots, p, \qquad (b)$$

and the matrix $\mathbf{B} = (b_{kj})$ is nonsingular. Moreover,

$$\max_{1\le k\le p} \|\mathbf{A}\mathbf{u}_k^*\| = O\Big(\big[\mathrm{tr}(\mathbf{A}'\mathbf{A})\big]^{\frac{1}{2}} \Big), \text{ for any matrix } \mathbf{A}, \qquad (c)$$

where $\mathbf{u}_k^* = (u_{1k}, \ldots, u_{nk})'$ and $||\cdot||$ denotes the Euclidean norm.

Asymptotics

Assumption 2

The functions $g(\cdot)$ and $h_j(\cdot)$ satisfy the Lipschitz condition of order ν, i.e., there exists a constant c such that

$$|f_j(s) - f_j(t)| \leq c|s-t|^{(\nu)}, \quad \text{for any } s,t \in [0,1], \quad j = 0,1,\ldots,p,$$

where $f_0(\cdot) = g(\cdot)$ and $f_j(\cdot) = h_j(\cdot),\ j = 1,\ldots,p.$

Asymptotics

Lemma

If Assumptions 1 and 2 are satisfied, and ε_i are independent with mean zero and constant variance σ^2 and $\mu_{3i} = E\varepsilon_i^3$ being uniformly bounded, then

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to_D N(0, \sigma^2 \mathbf{B}^{-1}) \quad \text{and} \quad \mathbf{B}'_{m_n}(t)\hat{\boldsymbol{\theta}} - g(t) = O_p(n^{-\frac{\nu}{2\nu+1}}),$$

where "$\to_D$" denotes convergence in distribution and $\mathbf{B}$ is defined in Assumption 1. Proof is similar to Ahmed et al (2007).

Asymptotic Properties

- To study the asymptotic quadratic risks of $\hat{\boldsymbol{\beta}}_1^{UR}$, $\hat{\boldsymbol{\beta}}_1^{R}$, $\hat{\boldsymbol{\beta}}_1^{S}$ and $\hat{\boldsymbol{\beta}}_1^{S+}$, we consider a sequence of local alternatives
- Under fixed alternatives all the estimators are asymptotically equivalent to $\hat{\boldsymbol{\beta}}_1$, while $\hat{\boldsymbol{\beta}}_1^{R}$ has unbounded risk.
- A sequence $\{K_n\}$ of local alternatives defined by

$$K_n : \boldsymbol{\beta}_{2(n)} = n^{-\frac{1}{2}}\boldsymbol{\omega}, \ \boldsymbol{\omega} \neq \mathbf{0} \text{ fixed} \tag{4}$$

Asymptotic Properties

Asymptotic Distirbutional Bias (ADB)

The asymptotic distributional bias (ADB) of an estimator $\boldsymbol{\delta}$ is defined as

$$\text{ADB}(\boldsymbol{\delta}) = \lim_{n\to\infty} E\left\{ n^{\frac{1}{2}}(\boldsymbol{\delta} - \boldsymbol{\beta}_1) \right\}.$$

Asymptotic Distirbutional Bias (ADB)

Suppose that Assumptions 1 and 2 hold. Under $\{K_n\}$, the ADB of the estimators are as follows:

$$
\begin{aligned}
\text{ADB}(\hat{\beta}_1^{UR}) &= \mathbf{0}, \\
\text{ADB}(\beta_1^{R}) &= -\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\omega, \\
\text{ADB}(\hat{\beta}_1^{S}) &= -(p_2-2)\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\omega E(\chi^{-2}_{p_2,\alpha};\Delta), \\
\text{ADB}(\hat{\beta}_1^{S+}) &= -\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\omega H_{p_2+2}(p_2-2;\Delta) - (p_2-2)\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\omega \\
&\qquad \left\{ E\left[\chi^{-2}_{p_2+2}(\Delta) I(\chi^2_{p_2+2}(\Delta) > p_2-2)\right]\right\}
\end{aligned}
$$

where $\mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix}$ with $\mathbf{B}$ defined in Assumption 1, $\Delta = (\omega'\mathbf{B}_{22.1}\omega)\sigma^{-2}$, $\mathbf{B}_{22.1} = \mathbf{B}_{22} - \mathbf{B}_{21}\mathbf{B}_{11}^{-1}\mathbf{B}_{12}$, $H_v(x;\Delta)$ denotes the noncentral chi-square distribution function with noncentrality parameter Δ and v degrees of freedom.

RISK PROPERTIES

Asymptotic Distirbutional Risk (ADR)

- Define a quadratic loss function using a positive definite matrix (p.d.m.) $\mathbf{M}$, namely,

$$\mathcal{L}(\boldsymbol{\delta}, \boldsymbol{\beta}_1) = n(\boldsymbol{\delta} - \boldsymbol{\beta}_1)'\mathbf{M}(\boldsymbol{\delta} - \boldsymbol{\beta}_1),$$

  $\boldsymbol{\delta}$ can be any one of estimator of $\hat{\boldsymbol{\beta}}_1$,
- We assume that the asymptotic distribution function of $\boldsymbol{\delta}$ under $\{K_n\}$ exists and is given by

$$F(\mathbf{x}) = \lim_{n\to\infty} P\{\sqrt{n}(\boldsymbol{\delta} - \boldsymbol{\beta}_1) \leq \mathbf{x}|K_n\}$$

- Then, the AQDR of $\boldsymbol{\delta}$ is defined as

$$R(\boldsymbol{\delta}, \mathbf{M}) = \text{tr}\left\{\mathbf{M}\int_{\mathcal{R}^{p_1}}\int \mathbf{x}\mathbf{x}'dF(\mathbf{x})\right\} = \text{tr}(\mathbf{M}V),$$

  where $\mathbf{V}$ is the dispersion matrix for the asymptotic distribution $F(\mathbf{x})$.

RISK PROPERTIES

Asymptotic Quadratic Distributional Risk (AQDR)

Suppose that assumptions 1 and 2 hold, then under $\{K_n\}$ the AQDR of the estimators are:

$$
\begin{aligned}
R(\hat{\boldsymbol{\beta}}_1^{UR}; \mathbf{M}) &= \sigma^2 \mathrm{tr}(\mathbf{M}\mathbf{B}_{11.2}^{-1}), \\
R(\boldsymbol{\beta}_1^{R}; \mathbf{M}) &= \sigma^2 \mathrm{tr}(\mathbf{M}\mathbf{B}_{11}^{-1}) + \boldsymbol{\omega}'\mathbf{B}_{21}\mathbf{B}_{11}^{-1}\mathbf{M}\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\boldsymbol{\omega}, \\
R(\hat{\boldsymbol{\beta}}_1^{S}; \mathbf{M}) &= \sigma^2 \left[\mathrm{tr}(\mathbf{M}\mathbf{B}_{11.2}^{-1}) - (p_2 - 2)\mathrm{tr}(\mathbf{B}_{21}\mathbf{B}_{11}^{-1}\mathbf{M}\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\mathbf{B}_{22.1}^{-1}) \right. \\
&\qquad \left. \cdot \left\{2E(\chi^2_{p_2,\alpha}(\Delta)) - (p_2 - 2)E(\chi^{-4}_{p_2+2}(\Delta))\right\}\right] \\
&\qquad + (p_2^2 - 4)\boldsymbol{\omega}'\mathbf{B}_{21}\mathbf{B}_{11}^{-1}\mathbf{M}\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\boldsymbol{\omega}E(\chi^{-4}_{p_2+4}(\Delta)),
\end{aligned}
$$

RISK PROPERTIES

$$
\begin{aligned}
& R(\hat{\boldsymbol{\beta}}^{S+};\mathbf{M}) \\
= \; & R(\hat{\boldsymbol{\beta}}^{S};\mathbf{M}) + (p_2-2)\mathrm{tr}(\mathbf{B}_{21}\mathbf{B}_{11}^{-1}\mathbf{M}\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\mathbf{B}_{22.1}^{-1})\left\{2E\left[\chi^{-2}_{p_2+2}(\Delta)I(\chi^2_{p_2+2}(\Delta)\le p_2-2)\right]\right. \\
& \left.-(p_2-2)E\left[\chi^{-4}_{p_2+2}(\Delta)I(\chi^2_{p_2+2}(\Delta)\le p_2-2)\right]\right\} - \sigma^2\mathrm{tr}(\mathbf{B}_{21}\mathbf{B}_{11}^{-1}\mathbf{M}\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\mathbf{B}_{22.1}^{-1}) \\
& \cdot H_{p_2+2}(p_2-2;\Delta) + \boldsymbol{\omega}'\mathbf{B}_{21}\mathbf{B}_{11}^{-1}\mathbf{M}\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\boldsymbol{\omega}\left[2H_{p_2+2}(p_2-2;\Delta) - H_{p_2+4}(p_2-2;\Delta)\right] \\
- \; & (p_2-2)\boldsymbol{\omega}\mathbf{B}_{21}\mathbf{B}_{11}^{-1}\mathbf{M}\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\boldsymbol{\omega}'\left\{2E\left[\chi^{-2}_{p_2+2}(\Delta)I(\chi^2_{p_2+2}(\Delta)\le p_2-2)\right]\right. \\
- \; & \left. 2E\left[\chi^{-2}_{p_2+2}(\Delta)I(\chi^2_{p_2+2}(\Delta)\le p_2-2)\right] + (p_2-2)E\left[\chi^{-4}_{p_2+2}(\Delta)I(\chi^2_{p_2+2}(\Delta)\le p_2-2)\right]\right\},
\end{aligned}
$$

THEORETICAL SUMMARY

Dominance of Shrinkage Estimator

By comparing $\mathrm{R}(\hat{\boldsymbol{\beta}}_1^S)$ and $\mathrm{R}(\hat{\boldsymbol{\beta}}_1^R)$ following dominance condition holds. If $\mathbf{M} \in \mathbf{M}^{\mathbf{D}}$, $\hat{\boldsymbol{\beta}}_1^S$ dominates $\hat{\boldsymbol{\beta}}_1$ for any $\boldsymbol{\omega}$ in the sense of AQDR, where

$$\mathbf{M}^{\mathbf{D}} = \left\{ \mathbf{M} : \frac{\mathrm{tr}(\mathbf{B}_{21}\mathbf{B}_{11}^{-1}\mathbf{M}\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\mathbf{B}_{22.1}^{-1})}{\mathrm{ch}_{\max}(\mathbf{B}_{21}\mathbf{B}_{11}^{-1}\mathbf{M}\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\mathbf{B}_{22.1}^{-1})} \geq \frac{p_2+2}{2} \right\}.$$

THEORETICAL SUMMARY

Dominance of Shrinkage Estimator

- $\hat{\boldsymbol{\beta}}^{S+}$ dominates $\hat{\boldsymbol{\beta}}^{S}$ for all the values of $\boldsymbol{\omega}$, with strict inequality holds for some $\boldsymbol{\omega}$.
- Risk of $\hat{\boldsymbol{\beta}}^{S+}$ is also smaller than the risk of $\hat{\boldsymbol{\beta}}_1$ in the entire parameter space and the upper limit is attained when Δ approaches ∞.
- This implies that

$$\mathrm{R}(\hat{\boldsymbol{\beta}}_1^{S+}) \leq \mathrm{R}(\hat{\boldsymbol{\beta}}_1^{S}) \leq \mathrm{R}(\hat{\boldsymbol{\beta}}_1), \text{ for any } \mathbf{M} \in \mathbf{M}^{\mathbf{D}} \text{ and } \boldsymbol{\omega},$$

with strict inequality holds for some $\boldsymbol{\omega}$.

Thus, we conclude that $\hat{\boldsymbol{\beta}}_1^{S}$ and $\hat{\boldsymbol{\beta}}_1^{S+}$ perform better than $\hat{\boldsymbol{\beta}}_1$ in the entire parameter space induced by Δ. The gain in risk over $\hat{\boldsymbol{\beta}}_1$ is substantial when $\Delta = 0$ or near.

MONTE CARLO RESULTS

Setup

We simulate the response from the following model:

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + \ldots, + x_{pi}\beta_p + g(t_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $t_i = (i - 0.5)/n$, $x_{1i} = (\zeta_{1i}^{(1)})^2 + \zeta_i^{(1)} + \xi_{1i}$,
$x_{2i} = (\zeta_{2i}^{(1)})^2 + \zeta_i^{(1)} + 2\xi_{2i}$, $x_{si} = (\zeta_{si}^{(1)})^2 + \zeta_i^{(1)}$ with
$\zeta_{si}^{(1)}$ i.i.d. $\sim N(0,1)$, $\zeta_i^{(1)}$ i.i.d. $\sim N(0,1)$,
ξ_{1i} $\sim$Bernoulli (0.45) and ξ_{2i} $\sim$Bernoulli (0.45) for all
$s = 3, \ldots, p$ and $i = 1, \ldots, n$.
Moreover, ε_i are i.i.d. $N(0,1)$, and $g(t) = \sin(4\pi t)$.

MONTE CARLO RESULTS

Hypotheses that we are testing

- $H_0 : \boldsymbol{\beta}_j = \mathbf{0}$, for $j = p_1 + 1, p_1 + 2, \ldots, p_1 + p_2$, with $p = p_1 + p_2$.
- We are interested to estimating β_1, β_2, β_3 and β_4 when the remaining regression parameters may not be useful.
- We partition the regression coefficients as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = (\boldsymbol{\beta}_1, \mathbf{0})$ with $\boldsymbol{\beta}_1 = (2, 1.5, 1, 0.6)$,
- We defined the parameter $\Delta^* = ||\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}||$, where $\boldsymbol{\beta}^{(0)} = (\boldsymbol{\beta}_1, \mathbf{0})$ and $||\cdot||$ is the Euclidean norm.
- To determine the behavior of the estimators for $\Delta^* > 0$, further datasets were generated from those distributions under local alternative hypotheses. We considered $\Delta^* = 0, .1, .2, .3, .4, .5, .8, 1, 2,$ and 4.

MONTE CARLO RESULTS

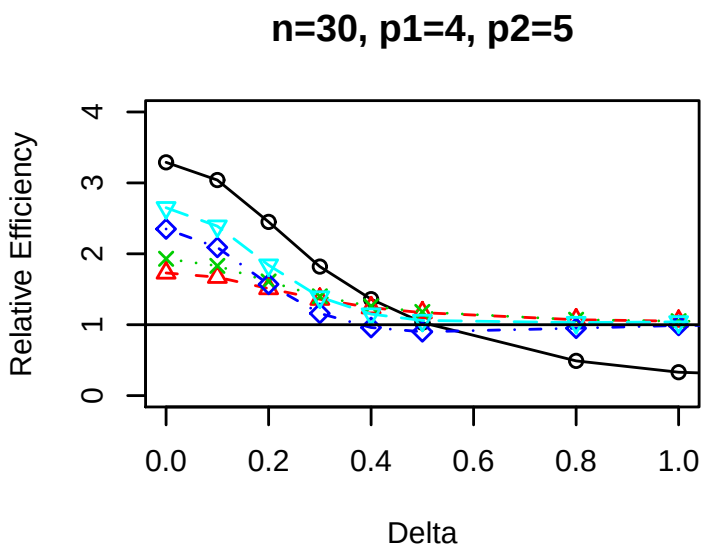
Relative MSE

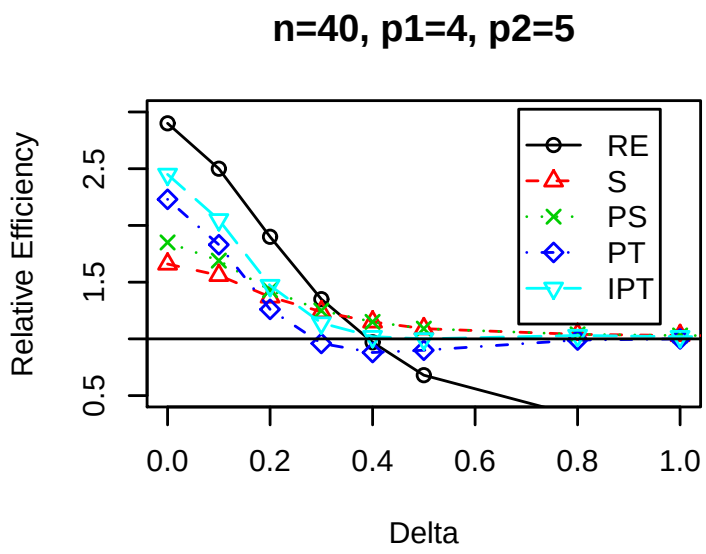
Numerically calculated the relative MSEs of the proposed estimators $\tilde{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_1^{\mathrm{S}}$, $\hat{\boldsymbol{\beta}}_1^{\mathrm{S+}}$, $\hat{\boldsymbol{\beta}}_1^{\mathrm{PT}}$, $\hat{\boldsymbol{\beta}}_1^{\mathrm{IPT}}$, relative to the unrestricted estimator $\hat{\boldsymbol{\beta}}_1$, using:

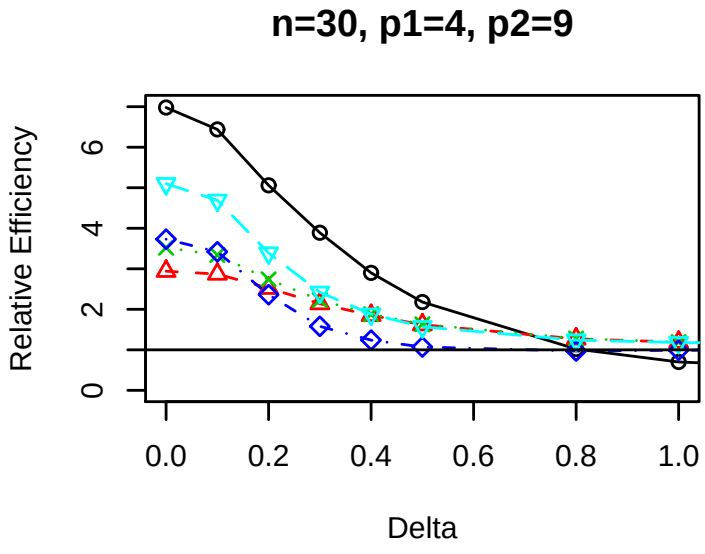
$$RMSE(\hat{\boldsymbol{\beta}}_1 : \hat{\boldsymbol{\beta}}_1^{\diamond}) = \frac{MSE(\hat{\boldsymbol{\beta}}_1)}{MSE(\hat{\boldsymbol{\beta}}_1^{\diamond})}.$$

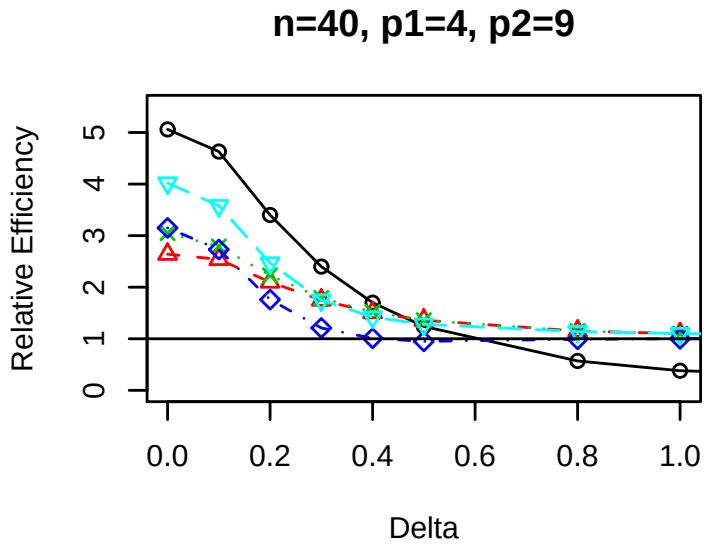
The amount by which an RMSE is larger than unity indicates the degree of superiority of the estimator $\hat{\boldsymbol{\beta}}_1^{\diamond}$ over $\hat{\boldsymbol{\beta}}_1$.

RMSEs

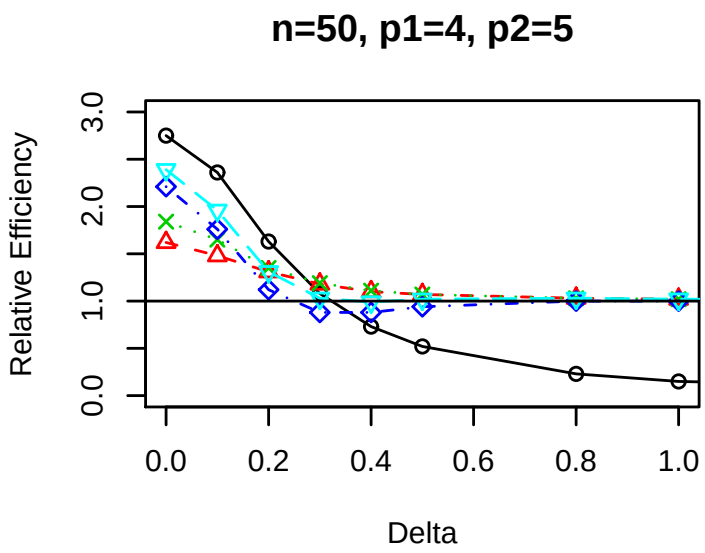

n=30, p1=4, p2=5
Relative Efficiency
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
Delta

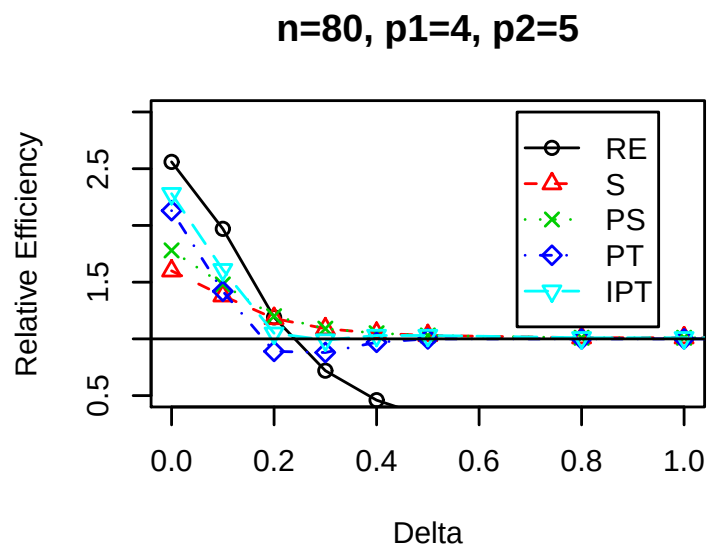

n=40, p1=4, p2=5
Relative Efficiency
0.5
1.5
2.5
0.0
0.2
0.4
0.6
0.8
1.0
Delta
RE
S
PS
PT
IPT

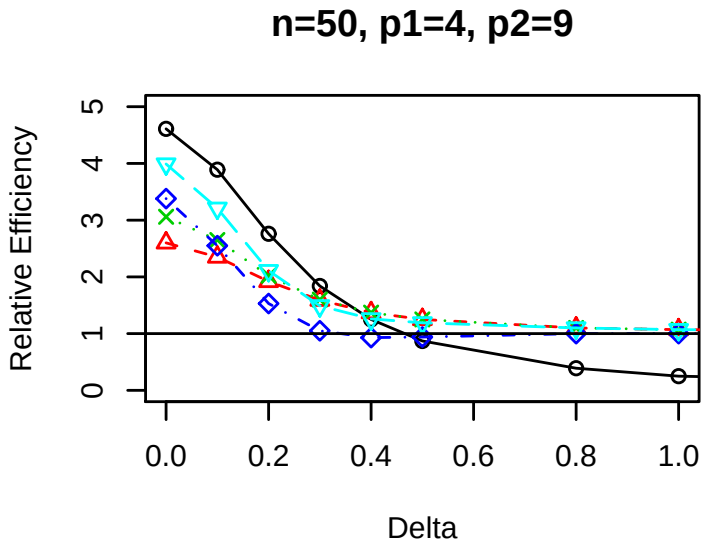

n=30, p1=4, p2=9
Relative Efficiency
0
2
4
6
0.0
0.2
0.4
0.6
0.8
1.0
Delta

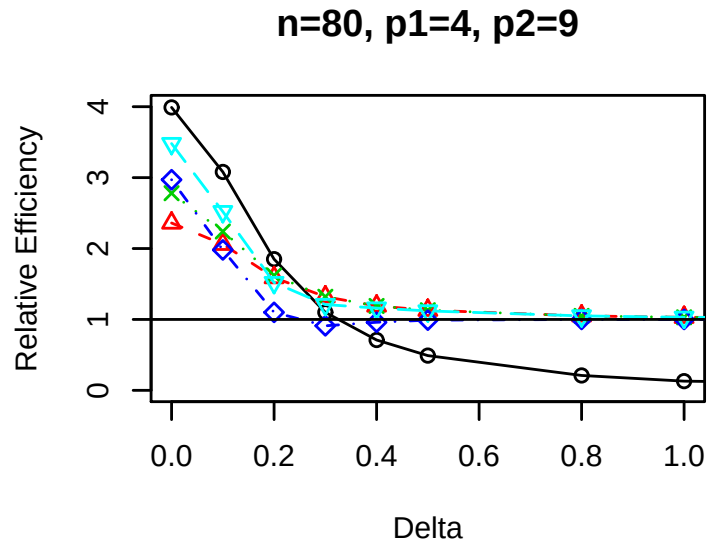

n=40, p1=4, p2=9
Relative Efficiency
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
Delta

RMSEs


n=50, p1=4, p2=5
Relative Efficiency
0.0
1.0
2.0
3.0
0.0
0.2
0.4
0.6
0.8
1.0
Delta


n=80, p1=4, p2=5
Relative Efficiency
0.5
1.5
2.5
0.0
0.2
0.4
0.6
0.8
1.0
Delta
RE
S
PS
PT
IPT


n=50, p1=4, p2=9
Relative Efficiency
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
Delta


n=80, p1=4, p2=9
Relative Efficiency
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
Delta

Shrinkage Vs. APE

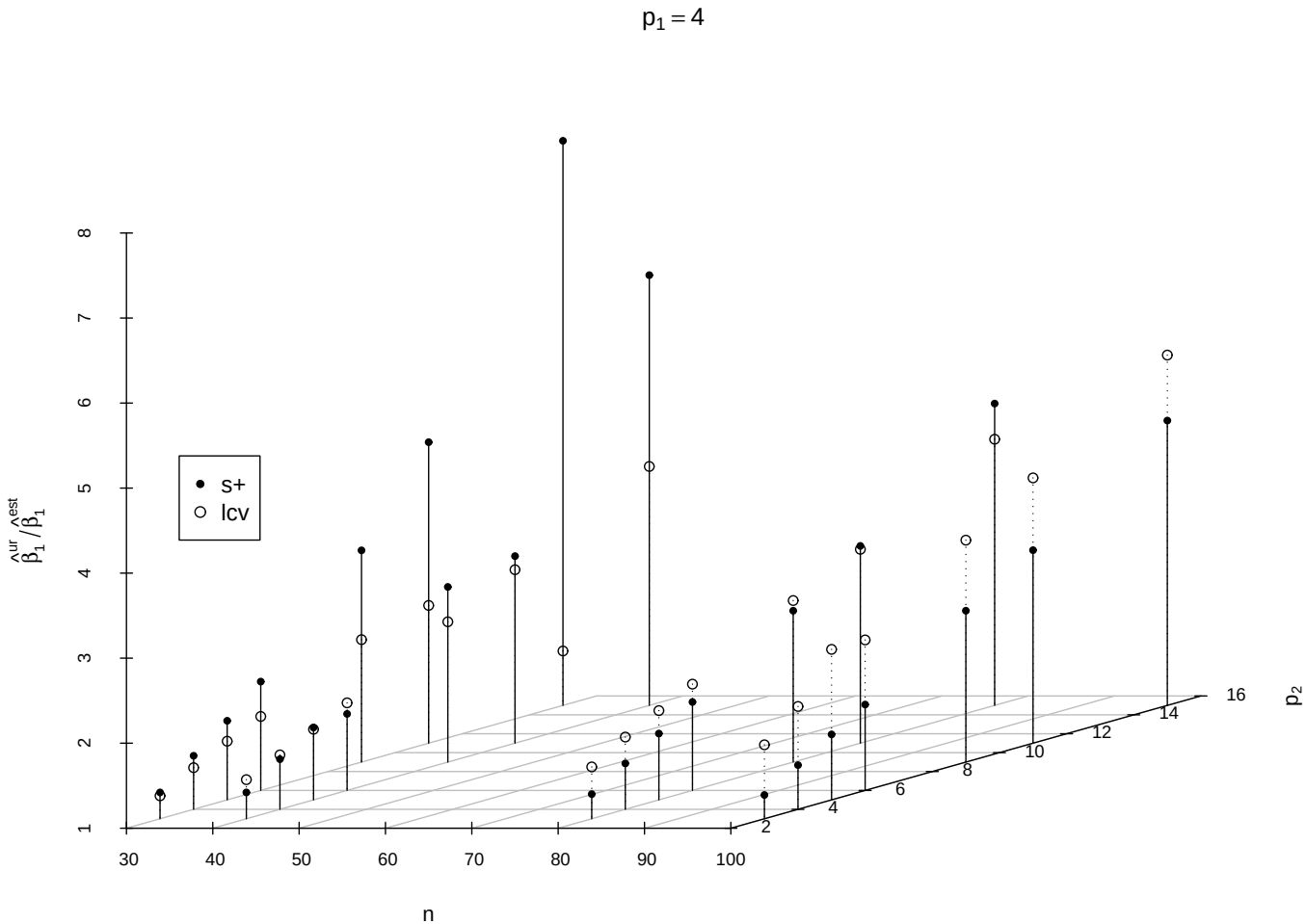


Figure: Three-dimensional plot of simulated relative MSE against n and p_2 to compare a positive shrinkage estimator and APE (CV) when $p_1 = 4$.

Summary

- Application to a real data shows the proposed method performs equally, in terms of loglikelihood values, with the absolute penalty estimator (our model has smaller number covariates)
- In Monte Carlo simulation, we found the positive-shrinkage estimator have smaller relative risk when the number of nuisance parameters are large.
- APE performs the best when p_1 and n are large

Future/Current work

- Shrinkage and APE estimation in multiple linear regression models
- Software for shrinkage estimation

`shrink`: R-package for Shrinkage Estimation

- Provides James-Stein-type shrinkage estimates of the regression parameters in a linear regression model.
- Computes unrestricted (ur), restricted (res), shrinkage (s), and positive shrinkage (ps) estimates of slope parameters.

Technically:

- `shrink()` produces an object of class "`shrink`" for which `coef()`, `predict()`, `residuals()` and `plot()` methods are available.
- Based on S3 methods of writing R extensions.

Usage: Options & Arguments

```
shrink(formula, formula2 = NULL, data = list(),
method = "AIC", est = "ps", test = "F", ...)
```

`formula`	A full model formula in the form of $y \sim x1 + x2$;
`formula2`	Optional sub-model in the form of $y \sim x1$; automatically selected if not supplied
`data`	Optional `data.frame`, or a matrix whose first column is the response
`method`	Sub-model selection method, either "AIC" or "BIC"
`est`	Estimators: one of "ur", "res", "s", "ps". Default is "ps"
`test`	Test statistic; either "Chisq" or "F"; default is "F"
`...`	Other arguments that can be passed to `shrink`

Example :

```
shrink(bodyfat ~., data=bodyfat, est=''ps'',
method=''BIC'', test=''Chisq'')
```

Data source: `http://lib.stat.cmu.edu/datasets/bodyfat`

Thank you!

References

Ahmed, S.E., Doksum, K. A., Hossain, S. and You, J. (2007)

Shrinkage, Pretest and Absolute Penalty Estimators in Partially linear models

Australian & New Zealand Journal of Statistics, 49, 435-454.

Engle, R.F. and W.J. Granger and J. Rice and A. Weiss (1986)

Semiparametric estimates of the relation between weather and electricity sales

Journal of the American Statistical Association, 80, 310 – 319

Fox, John (2002)

An R and S-PLUS Companion to Applied Regression

Sage Publications, Thousand Oaks

Fox, John (2005)

Introduction to Nonparametric Regression

http://socserv.socsci.mcmaster.ca/jfox/Courses/Oxford-2005/index.html

References

Long, J. S. (1997)

Regression models for categorical and limited dependent variables,

SAGE Publications, Thousand Oaks.

Mroz, T. A. (1987)

The sensitivity of an empirical model of married womens hours of work to economic and statistical assumptions,

Econometrica 55(4), 765799.

Wooldridge, J. (2003)

Introductory econometrics: a modern approach

South Western College Publishing.

Tibshirani, R. (1996).

Regression shrinkage and selection via the lasso.

J. Royal. Statist. Soc B., Vol. 58, No. 1, pages 267-288.

References

A.K.Md. E. Saleh (2006).

Theore of Preliminary Testand Stein-Type Estimation with Applications

Wiley.

Ahmed, S. E. , Saleh, E., Volodin , A. I. and I. N. Volodin (2007).

Asymptotic expansion of the coverage probability of James-Stein Estimators.

Journal of Theory of Probability and its Applications, 51, 683-695.