Robust Variable Selection with Exponential Squared Loss

Heping Zhang

Department of Biostatistics
Yale University School of Public Health

This talk is based on a joint work with
Xueqin Wang, Yunlu Jiang, and Mian Huang

June 28, 2013

Outline

Data and Model

- Assume that $\{(\mathbf{x}_i, Y_i), i = 1, \ldots, n\}$ is a random sample from population $(\mathbf{x}, Y)$.
- Let $D_i = (\mathbf{x}_i, Y_i)$ and $\mathbf{D}_n = (D_1, \cdots, D_n)$ be the observed data.
- Y is a univariate response.
- $\mathbf{x}$ is a d-dimensional predictor.
- $(\mathbf{x}, Y)$ has joint distribution F. And,

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{\beta}$ is a d-dimensional vector of unknown parameters, and the error terms ε_i are i.i.d. with unknown distribution G, $E(\varepsilon_i) = 0$, and ε_i is independent of $\mathbf{x}_i$.

The Goal

- Some of the coefficients in β are zero and their variables do not contribute to Y_i.
- Without loss of generality, let $\beta = (\beta_1^T, \beta_2^T)^T$, where $\beta_1 \in \mathbb{R}^s$ and $\beta_2 \in \mathbb{R}^{d-s}$.
  - ◇ The true regression coefficients are $\beta_0 = (\beta_{01}^T, \beta_{02}^T)^T$ with each element of β_{01} being nonzero, and $\beta_{02} = \mathbf{0}$.
- Let $\mathbf{x}_i = (\mathbf{x}_{i1}^T, \mathbf{x}_{i2}^T)^T$, where $\mathbf{x}_{i1}$ and $\mathbf{x}_{i2}$ are the covariates corresponding to β_1 and β_2.

⊙ Select important explanatory variables in the regression model.

Literature

- ∞ methods . . . $\rightsquigarrow$ penalization methods
- Still ∞ . . . examples include:
  - Bridge regression (Frank and Friedman, 1993)
  - LASSO (Tibshirani, 1996)
  - SCAD (Fan and Li, 2001)
  - Adaptive LASSO (Zou, 2006)

How to Deal with Outliers

- Many of the methods are closely related to the least squares method
  - $\Rightarrow$ The least squares method is sensitive to outliers with finite samples
  - $\hat{o}o$ Outliers can present serious problems

★ In the presence of outliers, how to replace the least squares criterion with a robust one?

Penalized Robust Regression

Fan and Li (2001) introduced a general framework of penalized robust regression estimators, i.e., to minimize

$$\Gamma_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \phi(Y_i - \mathbf{x}_i^T \boldsymbol{\beta}) + n \sum_{j=1}^{p} p_{\lambda_{nj}}(|\beta_j|) \tag{2}$$

with respect to $\boldsymbol{\beta}$, where $\phi(\cdot)$ is the loss function and $p_{\lambda_{nj}}(\cdot)$ the penalty function.

Penalized Robust Regression

The varieties include

- Wang et al. (2007): $\phi(t) = |t|$, $p_{\lambda_{nj}}(|\beta_j|) = \lambda_{nj}\,|\beta_j|$
- Wu and Liu (2009): the penalized quantile regression $\phi(t) = t\{\tau - \mathbf{1}(t < 0)\}$ with $0 \leq \tau \leq 1$, and $p_{\lambda_{nj}}(|\beta_j|)$ is either the SCAD penalty or the adaptive LASSO penalty
- Kai et al. (2011): a penalized composite quantile loss (Zou and Yuan, 2008)
- Johnson and Peng (2008): a rank-based approach
- Wang and Li (2009): a weighted Wilcoxon-type SCAD
- Leng (2010): regularized rank regression
- Bradic et al. (2011): the penalized composite quasi-likelihood

Basic Questions

- What is the breakdown point for a penalized robust regression estimator?
- Is its influence function bounded?

Motivation and Idea

- To study the robustness of variable selection procedures
  - $\Rightarrow$ We need to introduce a new robust variable selection procedure
  - $\Leftarrow$ We use the exponential loss function as in Adaboost (Friedman et al., 2000)

Outline

Exponential Squared Loss

Exponential squared loss is defined as

$$\phi_\gamma(t) = 1 - \exp(-t^2/\gamma),$$

where γ controls the degree of robustness for the estimators.

- When γ is large, $\phi_\gamma(t) \approx t^2/\gamma \Rightarrow$ the least squares.
- For a small γ, observations with large absolute values of $t_i = Y_i - \mathbf{x}_i^T\beta$ will result in large losses of $\phi_\gamma(t_i)$, and therefore have a small impact on the estimation of β.
  - $\Uparrow$ A smaller γ would limit the influence of an outlier on the estimators.
  - $\Downarrow$ It could also reduce the sensitivity of the estimators.

★ How to select γ so that the estimators are robust and possess desirable finite and large sample properties?

Adaptive Loss

We maximize

$$\ell_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \exp\{-(Y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2/\gamma_n\} - n\sum_{j=1}^{d} p_{\lambda_{nj}}(|\beta_j|) \tag{3}$$

with respect to $\boldsymbol{\beta}$.

- This is a special case of (2) for any given $\gamma_n \in (0, +\infty)$.
- (3) is distinct from (2) because γ_n is adaptive and data driven.
  $\Rightarrow$ A high breakdown point and high efficiency.

Outline

Notation

Let $\hat{\boldsymbol{\beta}}_n = (\hat{\boldsymbol{\beta}}_{n1}^T, \hat{\boldsymbol{\beta}}_{n2}^T)^T$ be the resulting estimator of (3),

$$a_n = \max\{p'_{\lambda_{nj}}(|\beta_{0j}|) : \beta_{0j} \neq 0\},$$

$$b_n = \max\{p''_{\lambda_{nj}}(|\beta_{0j}|) : \beta_{0j} \neq 0\},$$

and

$$I(\boldsymbol{\beta}, \gamma) = \frac{2}{\gamma} \int \mathbf{x}\mathbf{x}^T e^{-r^2/\gamma} \left(\frac{2r^2}{\gamma} - 1 \right) dF(\mathbf{x}, y), \; r = Y - \mathbf{x}^T\boldsymbol{\beta}.$$

Regularity Condition

Assume

(C1): $\Sigma = E(\mathbf{x}\mathbf{x}^T)$ is positive definite, and $E||\mathbf{x}||^3 < \infty$.

- Condition (C1) ensures that the main term dominates the remainder in the Taylor expansion.

Consistency and Normality

Theorem

Assume that condition (C1) holds, $b_n = o_p(1)$, and $I(\beta_0, \gamma_0)$ is negative definite.
(i) If $\gamma_n - \gamma_0 = o_p(1)$ for some $\gamma_0 > 0$, there exists a local maximizer $\hat{\beta}_n$ such that $\|\hat{\beta}_n - \beta_0\| = O_p(n^{-1/2} + a_n)$.
(ii) (Oracle property) If $\sqrt{n}a_n = O_p(1)$, $1/\min_{s+1\le j\le d}(\sqrt{n}\lambda_{nj}) = o_p(1)$, $\sqrt{n}(\gamma_n - \gamma_0) = o_p(1)$, and with probability 1,

$$\lim_{n\to\infty} \inf \lim_{t\to 0+} \inf \left\{ \min_{s+1\le j\le d} p'_{\lambda_{nj}}(|t|)/\lambda_{nj} \right\} > 0, \tag{4}$$

then we have (a) sparsity, i.e., $\hat{\beta}_{n2} = \mathbf{0}$ with probability 1 and (b) asymptotic normality.

Asymptotic Normality

Specifically,

$$\sqrt{n}(I_1(\boldsymbol{\beta}_{01},\gamma_0)+\Sigma_1)\left\{\hat{\boldsymbol{\beta}}_{n1}-\boldsymbol{\beta}_{01}+(I_1(\boldsymbol{\beta}_{01},\gamma_0)+\Sigma_1)^{-1}\Delta\right\} \tag{5}$$
$$\xrightarrow{D} N\left(\mathbf{0},\Sigma_2\right),$$

where

$$\Sigma_1 = \text{diag}\{p''_{\lambda_{n1}}(|\beta_{01}|),\cdots,p''_{\lambda_{ns}}(|\beta_{0s}|)\},$$
$$\Sigma_2 = cov(\exp(-r^2/\gamma_0)\frac{2r}{\gamma_0}\mathbf{x}_{i1}),$$
$$\Delta = \left(p'_{\lambda_{n1}}(|\beta_{01}|)\text{sign}(\beta_{01}),\cdots,p'_{\lambda_{ns}}(|\beta_{0s}|)\text{sign}(\beta_{0s})\right)^T,$$
$$I_1(\boldsymbol{\beta}_{01},\gamma_0) = \frac{2}{\gamma_0}E\left[\exp\left(-r^2/\gamma_0\right)\left(\frac{2r^2}{\gamma_0}-1\right)\right]\left(E\mathbf{x}_{i1}\mathbf{x}_{i1}^T\right).$$

Special Cases

- Some penalties do not satisfy the conditions in the theorem.
  - LASSO is inconsistent, and the oracle property does not hold.
  - Zou (2006) proposed the adaptive LASSO, and showed that it enjoys the oracle property.
- The adaptive LASSO penalty: $p_{\lambda_{nj}}(|\beta_j|) = \lambda_{nj}|\beta_j|$, $\lambda_{nj} = \tau_{nj}/|\tilde{\beta}_j|^k$ for some $k > 0$, where $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \cdots, \tilde{\beta}_d)^T$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}_0$, and τ_{nj}'s are the regularization parameters.
- With this penalty in (3), the estimators are $\sqrt{n}$-consistent and have the oracle property under the following condition (C2) in addition to the regularity condition (C1).
- (C2): $\max_{1\le j\le s}(\sqrt{n}\lambda_{nj}) = o_p(1)$ and $1/\min_{s+1\le j\le d}(\sqrt{n}\lambda_{nj}) = o_p(1)$.

Special Cases

- $\max_{1\le j\le s}(\sqrt{n}\lambda_{nj}) = o_p(1)$ implies $\sqrt{n}a_n = O_p(1)$ by the definition of a_n.
- Some data-driven methods for selecting λ_{nj}, e.g., cross-validation, may not satisfy condition (C2).
- (C2) holds for BIC.

Outline

Finite Sample Breakdown Point

⊙ To measure the maximum fraction of outliers in a sample that an estimator can tolerate before returning arbitrary values (Hampel, 1971; Donoho, 1982; Donoho and Huber, 1983).

- $\mathbf{D}_m = \{D_1, \cdots, D_m\}$ m bad points.
- $\mathbf{D}_{n-m} = \{D_{m+1}, \cdots, D_n\}$ $n-m$ good points.
- Let $\hat{\beta}(\mathbf{D}_n)$ denote a regression estimator based on sample $\mathbf{D}_n$.
- The addition breakdown point of an estimator $\hat{\beta}_n$:

$$BP(\hat{\beta}_n; \mathbf{D}_{n-m}) = \min \left\{ \frac{m}{n} : \sup_{\mathbf{D}_m} \|\hat{\beta}(\mathbf{D}_n) - \hat{\beta}(\mathbf{D}_{n-m})\| = \infty \right\},$$

where $||\cdot||$ is the Euclidean norm.

Note: The notation $\mathbf{D}_n$ is somewhat abused here for convenience. In the regression setting, many estimators such as S-estimator (Rousseeuw and Yohai, 1984), MM-estimator, τ-estimator (Yohai and Zamar, 1988), and REWLS-estimator (Gervini and Yohai, 2002), can achieve the highest asymptotic breakdown point of $1/2$.

Finite Sample Breakdown Point

What is the breakdown point, denoted by $BP(\hat{\boldsymbol{\beta}}_n; \mathbf{D}_{n-m}, \gamma_n)$, for our $\hat{\boldsymbol{\beta}}_n$ with the tuning parameter γ_n?

- Take an initial estimator $\tilde{\boldsymbol{\beta}}_n$.
- $$\zeta(\gamma_n) = \frac{2m}{n} + \frac{2}{n} \sum_{i=m+1}^{n} \phi_{\gamma_n} \left\{ r_i(\tilde{\boldsymbol{\beta}}_n) \right\},$$

  where $r_i(\boldsymbol{\beta}) = Y_i - \mathbf{x}_i^T \boldsymbol{\beta}$. Note that $\zeta(\gamma_n) \in (0, 2]$.
- $$a_{nm} = (n-m)^{-1} \max_{\boldsymbol{\beta} \in \mathbb{R}^d} \sharp\{i : m+1 \le i \le n \text{ and } \mathbf{x}_i^T \boldsymbol{\beta} = 0\}.$$
- If a set of d regressor variables is linearly independent, then $a_{nm} = (d-1)/(n-m)$.

Finite Sample Breakdown Point

Theorem

For any penalty function of the form $p_{\lambda_{nj}}(|\beta_j|) = \lambda_{nj} g(|\beta_j|)$, *where* $g(\cdot)$ *is a strictly increasing and unbounded function defined on* $[0, \infty]$, *and the weight* λ_{nj} *is positive for all* $j = 1, \cdots, d$. *If* $m/n \leq \epsilon < (1 - 2a_{nm})/(2 - 2a_{nm})$, $a_{nm} < 0.5$, *and* $\zeta(\gamma_n) < (1 - \epsilon)(2 - 2a_{nm})$ *hold, then, for any initial estimator* $\tilde{\boldsymbol{\beta}}_n$ *of* $\boldsymbol{\beta}_0$, *we have*

$$BP(\hat{\boldsymbol{\beta}}_n; \mathbf{D}_{n-m}, \gamma_n) \geq \min \left\{ BP(\tilde{\boldsymbol{\beta}}_n; \mathbf{D}_{n-m}), \frac{1 - 2a_{nm}}{2 - 2a_{nm}}, 1 - \frac{\zeta(\gamma_n)}{2 - 2a_{nm}} \right\}.$$

Finite Sample Breakdown Point

What does the theorem tell?

- It provides the lower bound for the breakdown point.
- The bound depends on the breakdown point of an initial estimate and the tuning parameter γ_n.
- If $\tilde{\beta}_n$ is a robust estimator with asymptotic breakdown point $1/2$, and γ_n is chosen such that $\zeta(\gamma_n) \in (0, 1]$, then $BP(\hat{\beta}_n; \mathbf{D}_{n-m}, \gamma_n)$ is asymptotically $1/2$.

$\Rightarrow$ How to select γ_n?

Finite Sample Breakdown Point

What penalties can the theorem be applied for?

+ LASSO, adaptive LASSO, the L_q penalty with $q > 0$, logarithm penalty, elastic-net penalty, and adaptive elastic-net penalty.
? SCAD (Fan and Li, 2001) and MCP (Zhang, 2010).

Outline

Influence Function

⊙ To measure the stability of estimators given an infinitesimal contamination (Hampel, 1968).

- $\delta_{\mathbf{z}}$: the point mass probability distribution at a fixed point $\mathbf{z} = (\mathbf{x}_0, y_0)^T \in \mathbb{R}^{d+1}$.
- Given the distribution F of $(\mathbf{x}, Y)$ in $\mathbb{R}^{d+1}$ and proportion $\epsilon \in (0, 1)$, the mixture distribution of F and $\delta_{\mathbf{z}}$ is $F_\epsilon = (1 - \epsilon)F + \epsilon\delta_{\mathbf{z}}$.

Influence Function

Suppose that λ_{nj}'s have the limit point λ_{0j}'s. Define

$$\beta_0^* = \arg\min_{\beta} \left[\left\{ \int \left(1 - e^{-(y-\mathbf{x}^T\beta)^2/\gamma_0}\right) dF \right\} + \sum_{j=1}^{d} p_{\lambda_{0j}}(|\beta_j|) \right],$$

$$\beta_\epsilon^* = \arg\min_{\beta} \left[\left\{ \int \left(1 - e^{-(y-\mathbf{x}^T\beta)^2/\gamma_0}\right) dF_\epsilon \right\} + \sum_{j=1}^{d} p_{\lambda_{0j}}(|\beta_j|) \right].$$

Note: β_0^* is a shrinkage of the true coefficient β_0 to 0.

- Let $IF_j(\mathbf{z}; \beta_0^*)$ be the j-th element of the influence function.

Influence Function

Theorem

$$IF_j(\mathbf{z};\beta_0^*) = \begin{cases} 0 & \text{if } \beta_{0j}^* = 0, \\ -\Gamma_j \left\{ 2\exp\left(-r_0^2/\gamma_0\right) r_0 \mathbf{x}_0/\gamma_0 + \nu_2 \right\}, & \text{otherwise,} \end{cases}$$

where Γ_j *denotes the* j*-th row of* $\{2A(\gamma_0)/\gamma_0 - B\}^{-1}$, $r_0 = y_0 - \mathbf{x}_0^T\beta_0^*$,

$$\nu_2 = \left\{ p'_{\lambda_{01}}(|\beta_{01}^*|)\,sign(\beta_{01}^*), \cdots, p'_{\lambda_{0d}}(|\beta_{0d}^*|)\,sign(\beta_{0d}^*) \right\}^T,$$

$$B = \mathrm{diag}\left\{ p''_{\lambda_{01}}(|\beta_{01}^*|), \cdots, p''_{\lambda_{0d}}(|\beta_{0d}^*|) \right\},$$

and

$$A(\gamma) = \int \mathbf{x}\mathbf{x}^T \exp\left\{-(y - \mathbf{x}^T\beta_0^*)^2/\gamma\right\} \left\{ \frac{2(y - \mathbf{x}^T\beta_0^*)^2}{\gamma} - 1 \right\} dF(\mathbf{x}, y).$$

Adaptive LASSO Penalty

If the regularization parameter is selected by the BIC described, according to condition (C2), we have $\lambda_{0j} = 0$ for $j = 1, \cdots, s$, and $\lambda_{0j} = +\infty$ for $j = s+1, \cdots, d$.

- The corresponding influence functions of the zero coefficients are zero.
- For the nonzero coefficients, the influence functions have the form

$$IF_j(\mathbf{z}; \boldsymbol{\beta}_0^*) = -\Gamma_j \left\{ 2 \exp\left(-r_0^2/\gamma_0\right) r_0 \mathbf{x}_{01}/\gamma_0 \right\}.$$

Outline

Regularization Parameter λ_{nj}

To reduce computational complexity and guarantee consistent variable selection, we choose the regularization parameter by minimizing a BIC-type objective function (Wang et al., 2007):

$$\sum_{i=1}^{n}\left[1-\exp\{-(Y_i-\mathbf{x}_i^T\boldsymbol{\beta})^2/\gamma_n\}\right]+n\sum_{j=1}^{d}\tau_{nj}\left|\beta_j\right|/|\tilde{\beta}_{nj}|-\sum_{j=1}^{d}\log(0.5n\tau_{nj})\log(n).$$

This leads to $\lambda_{nj}=\hat{\tau}_{nj}/|\tilde{\beta}_{nj}|$, where $\hat{\tau}_{nj}=\frac{\log(n)}{n}$.

Tuning Parameter γ_n

The tuning parameter γ_n controls the degree of robustness and efficiency of the proposed robust regression estimators.

- To select γ_n, we propose a data-driven procedure which yields both high robustness and high efficiency simultaneously.
- We determine a set of the tuning parameters such that the proposed penalized robust estimators have asymptotic breakdown point at $1/2$.
- We select the tuning parameter with the maximum efficiency.

Tuning Parameter γ_n

1. Find the pseudo outlier set of the sample: $S_n = 1.4826 \times median_i \left| r_i(\hat{\boldsymbol{\beta}}_n) - median_j(r_j(\hat{\boldsymbol{\beta}}_n)) \right|$. Then, take the pseudo outlier set $\mathbf{D}_m = \{(\mathbf{x}_i, Y_i) : |r_i(\hat{\boldsymbol{\beta}}_n)| \geq 2.5S_n\}$, set $m = \sharp\{1 \leq i \leq n : |r_i(\hat{\boldsymbol{\beta}}_n)| \geq 2.5S_n\}$, and $\mathbf{D}_{n-m} = \mathbf{D}_n / \mathbf{D}_m$.
2. Update the tuning parameter γ_n: Let γ_n be the minimizer of $\det(\hat{V}(\gamma))$ in the set $G = \{\gamma : \zeta(\gamma) \in (0, 1]\}$, where $\hat{V}(\gamma) = \{\hat{I}_1(\hat{\boldsymbol{\beta}}_n)\}^{-1} \tilde{\Sigma}_2 \{\hat{I}_1(\hat{\boldsymbol{\beta}}_n)\}^{-1}$, and

$$
\begin{aligned}
\hat{I}_1(\hat{\boldsymbol{\beta}}_n) &= \frac{2}{\gamma} \left\{ \frac{1}{n} \sum_{i=1}^{n} \exp\left(-r_i^2(\hat{\boldsymbol{\beta}}_n)/\gamma\right) \left(\frac{2r_i^2(\hat{\boldsymbol{\beta}}_n)}{\gamma} - 1 \right) \right\} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \right), \\
\tilde{\Sigma}_2 &= cov\left\{ \exp\left(-r_1^2(\hat{\boldsymbol{\beta}}_n)/\gamma\right) \frac{2r_1(\hat{\boldsymbol{\beta}}_n)}{\gamma} \mathbf{x}_1, \cdots, \exp\left(-r_n^2(\hat{\boldsymbol{\beta}}_n)/\gamma\right) \frac{2r_n(\hat{\boldsymbol{\beta}}_n)}{\gamma} \mathbf{x}_n \right\} \\
\zeta(\gamma_n) &= \frac{2m}{n} + \frac{2}{n} \sum_{i=m+1}^{n} \phi_{\gamma_n} \left\{ r_i(\tilde{\boldsymbol{\beta}}_n) \right\}.
\end{aligned}
$$

3. Update $\hat{\boldsymbol{\beta}}_n$.

Implementation

- Set $\hat{\boldsymbol{\beta}}_n$ as the MM estimator and detect the outliers in Step 1.
- Compute $\zeta(\gamma_n)$.
- By the breakdown theorem, asymptotic breakdown point is $1/2$.
- To attain a high efficiency, choose the tuning parameter γ_n by minimizing the determinant of asymptotic covariance matrix in Step 2.
- Since the calculation of of $\det(\hat{V}(\gamma))$ depends on estimate $\hat{\boldsymbol{\beta}}_n$, update $\hat{\boldsymbol{\beta}}_n$ in Step 3 using the block coordinate gradient descent (BCGD) algorithm (Tseng and Yun, 2009).
- Repeat Steps 1-3 once.

Outline

Design

- $n = 100, 200, 400, 800$, $d = 8$, and $\boldsymbol{\beta} = (1, 1.5, 2, 1, 0, 0, 0, 0)^T$.
- Generate $\mathbf{x}_i = (x_{i1}, \cdots, x_{id})^T$ from a multi-normal distribution $N(\mathbf{0}, \Omega_2)$, where the (i, j)-th element of Ω_2 is $\rho^{|i-j|}$, $\rho = 0.5$.
- The error term follows a Cauchy distribution.
- Replicate 1000 times.

Settings of Influential Points

- Influential points in the predictors: Covariate $\mathbf{x}_i$ follows a mixture of d-dimensional normal distributions $0.8N(\mathbf{0}, \Omega_1) + 0.2N(\boldsymbol{\mu}, \Omega_2)$, $\Omega_1 = I_{d\times d}$, $\boldsymbol{\mu} = 3\mathbf{1}_d$, $\mathbf{1}_d$ is d-dimensional vector of ones, and the error term follows a standard normal distribution;
- Influential points in the response: Covariate $\mathbf{x}_i$ follows a multi-normal distribution $N(\mathbf{0}, \Omega_2)$, and the error term follows a mixture normal distribution $0.8N(0, 1) + 0.2N(10, 6^2)$;
- Influential points in both the predictors and response: Covariate $\mathbf{x}_i$ follows a mixture of d-dimensional normal distributions $0.8N(\mathbf{0}, \Omega_1) + 0.2N(\boldsymbol{\mu}, \Omega_2)$, and the error term follows a Cauchy distribution.

Competitive Methods

- CQR-LASSO: the composite quantile regression (Zou and Yuan, 2008).
  - ◇ We set the quantiles $\tau_k = k/10$ for $k = 1, 2, \cdots, 9$.
- LAD-LASSO: least absolute deviation (Wang et al., 2007).
- The oracle method based on MM-estimator.
- Our method (ESL-LASSO).

Performance Measures

The performance is compared by

- the positive selection rate (PSR) – the proportion of selected causal features related to all causal features (Chen and Chen, 2008);
- the non-causal selection rate (NSR) – the average restricted only to the true zero coefficients (Fan and Li, 2001);
- and the median and median absolute deviation (MAD) of the model error (Fan and Li, 2001).
- the model error

$$ME = \left(\hat{\beta}_n - \beta_0\right)^T E\left[\mathbf{x}\mathbf{x}^T\right]\left(\hat{\beta}_n - \beta_0\right).$$

Result: Setting 1

n	Method	$\bar{\gamma}_n$	$\bar{\zeta}(\gamma_n)$	PSR	NSR	Model error	
						Median	MAD
100	ESL-LASSO	3.965	0.260	0.982	0.999	0.076	0.040
	CQR-LASSO	−−	−−	1.000	0.877	0.041	0.021
	LAD-LASSO	−−	−−	1.000	0.581	0.057	0.030
	Oracle	−−	−−	1.000	1.000	0.034	0.018
200	ESL-LASSO	4.450	0.309	1.000	1.000	0.027	0.013
	CQR-LASSO	−−	−−	1.000	0.935	0.019	0.010
	LAD-LASSO	−−	−−	1.000	0.539	0.028	0.012
	Oracle	−−	−−	1.000	1.000	0.017	0.009
400	ESL-LASSO	4.500	0.331	1.000	1.000	0.012	0.006
	CQR-LASSO	−−	−−	1.000	0.966	0.010	0.005
	LAD-LASSO	−−	−−	1.000	0.498	0.0142	0.007
	Oracle	−−	−−	1.000	1.000	0.009	0.005
800	ESL-LASSO	4.500	0.338	1.000	1.000	0.005	0.003
	CQR-LASSO	−−	−−	1.000	0.988	0.005	0.002
	LAD-LASSO	−−	−−	1.000	0.498	0.007	0.003
	Oracle	−−	−−	1.000	1.000	0.004	0.002

Result: Setting 2

n	Method	$\bar{\gamma}_n$	$\bar{\zeta}(\gamma_n)$	PSR	NSR	Model error	
						Median	MAD
100	ESL-LASSO	4.315	0.454	0.939	1.000	0.352	0.231
	CQR-LASSO	−−	−−	1.000	0.781	0.066	0.033
	LAD-LASSO	−−	−−	1.000	0.738	0.113	0.061
	Oracle	−−	−−	1.000	1.000	0.051	0.026
200	ESL-LASSO	4.449	0.633	1.000	1.000	0.080	0.039
	CQR-LASSO	−−	−−	1.000	0.789	0.046	0.023
	LAD-LASSO	−−	−−	1.000	0.712	0.050	0.026
	Oracle	−−	−−	1.000	1.000	0.025	0.013
400	ESL-LASSO	4.496	0.638	1.000	1.000	0.027	0.012
	CQR-LASSO	−−	−−	1.000	0.864	0.021	0.010
	LAD-LASSO	−−	−−	1.000	0.686	0.023	0.011
	Oracle	−−	−−	1.000	1.000	0.012	0.006
800	ESL-LASSO	4.499	0.642	1.000	1.000	0.009	0.005
	CQR-LASSO	−−	−−	1.000	0.910	0.010	0.006
	LAD-LASSO	−−	−−	1.000	0.633	0.011	0.005
	Oracle	−−	−−	1.000	1.000	0.006	0.003

Result: Setting 3

n	Method	$\bar{\gamma}_n$	$\bar{\zeta}(\gamma_n)$	PSR	NSR	Model error	
						Median	MAD
100	ESL-LASSO	4.565	0.662	1.000	0.989	0.174	0.101
	CQR-LASSO	−−	−−	1.000	0.675	0.173	0.097
	LAD-LASSO	−−	−−	1.000	0.488	0.113	0.061
	Oracle	−−	−−	1.000	1.000	0.098	0.050
200	ESL-LASSO	3.654	0.722	1.000	1.000	0.058	0.030
	CQR-LASSO	−−	−−	1.000	0.778	0.068	0.031
	LAD-LASSO	−−	−−	1.000	0.487	0.051	0.025
	Oracle	−−	−−	1.000	1.000	0.049	0.023
400	ESL-LASSO	3.589	0.850	1.000	1.000	0.022	0.012
	CQR-LASSO	−−	−−	1.000	0.856	0.031	0.016
	LAD-LASSO	−−	−−	1.000	0.459	0.023	0.011
	Oracle	−−	−−	1.000	1.000	0.023	0.011
800	ESL-LASSO	3.525	0.833	1.000	1.000	0.010	0.005
	CQR-LASSO	−−	−−	1.000	0.912	0.015	0.008
	LAD-LASSO	−−	−−	1.000	0.435	0.011	0.005
	Oracle	−−	−−	1.000	1.000	0.012	0.006

Observations

- ESL-LASSO yields larger model error than LAD-LASSO and CQR-LASSO when the sample size is small, because it involves some consistent estimators in its selection procedure.
- As the sample size increases, the medians and MADs of the model error decrease in all three settings.
- Although ESL-LASSO's are always larger than CQR-LASSO's, they are smaller than LAD-LASSO's if the sample size is at least 200 in the first setting.
- ESL-LASSO's are smaller than both LAD-LASSO's and CQR-LASSO's if the sample size is large enough in Settings 2 and 3.
- The PSR is around 1 for all three methods in all settings.
- The NSR of the ESL-LASSO estimator is as close 1 while the NSR of the LAD-LASSO and CQR-LASSO ranges from 0.431 to 0.738, and from 0.675 to 0.988, respectively.

Outline

Boston Housing Price

The dataset was downloaded from `http://lib.stat.cmu.edu/datasets/boston` (Harrison and Rubinfeld, 1978; Belsley et al., 1980).
There are 506 observations in the dataset. The response variable is medv (median value of owner-occupied homes in thousand dollars), and there are 13 predictors.

Boston Housing Price

The predictors: crim (per capita crime rate by town), zn (proportion of residential land zoned for lots over 25,000 sq.ft), indus (proportion of non-retail business acres per town), chas (Charles River dummy variable: equal to 1 if tract bounds river; 0 otherwise), nox (nitrogen oxides concentration: parts per 10 million), rm (average number of rooms per dwelling), age (proportion of owner-occupied units built prior to 1940), dis (weighted mean of distances to five Boston employment centres), rad (index of accessibility to radial highways), tax (full-value property-tax rate per ten thousand dollar), ptratio (pupil-teacher ratio by town), black $(1000(Bk - 0.63)^2$, where Bk is the proportion of blacks by town), lstat (lower status of the population (percent)).

Boston Housing Price

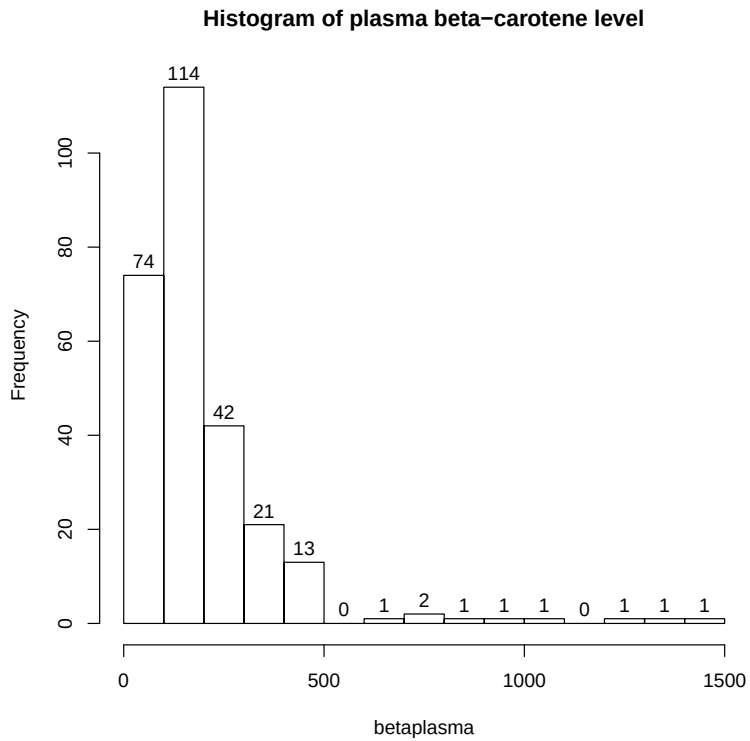
	Method				
Variable	ESL-LASSO	CQR-LASSO	LAD-LASSO	MM	OLS
crim	0	0	0	-0.097	-0.101
zn	0	0	0	0.072	0.118
indus	0	0	0	-0.005	0.015
chas	0	0	0	0.038	0.074
nox	0	0	0	-0.097	-0.224
rm	0.590	0.422	0.503	0.491	0.291
age	0	0	0	-0.117	0.002
dis	0	-0.057	-0.013	-0.235	-0.338
rad	0	0	0	0.156	0.290
tax	-0.105	-0.133	-0.058	-0.208	-0.226
ptratio	-0.076	-0.153	-0.155	-0.179	-0.224
black	0	0.040	0.085	0.124	0.092
lstat	-0.131	-0.334	-0.243	-0.174	-0.408

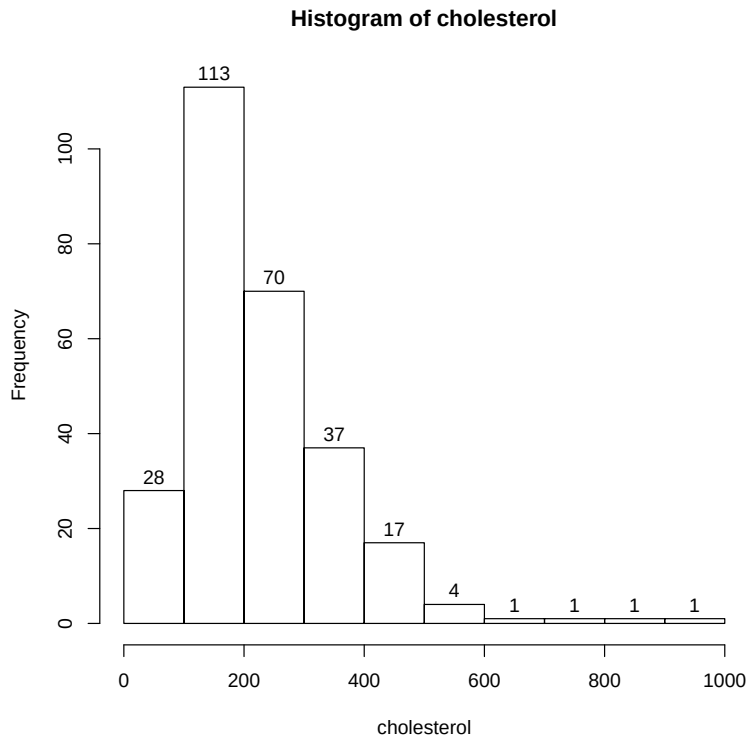
Outline

Plasma Beta-carotene Level

The dataset was downloaded from
`http://lib.stat.cmu.edu/datasets/Plasma_Retinol`.
We only analyze the data from the 273 female patients.
To study the relationships between the plasma beta-carotene level (betaplasma) and the following 10 covariates: age, smoking status (smokstat), quetelet, vitamin use (vituse), number of calories consumed per day (calories), grams of fat consumed per day (fat), grams of fiber consumed per day (fiber), number of alcoholic drinks consumed per week (alcohol), cholesterol consumed (cholesterol), and dietary beta-carotene consumed (betadiet).

Histograms of Betaplasma and Cholesterol





Histograms of Betaplasma and Cholesterol

	Method		
Variable	ESL-LASSO	CQR-LASSO	LAD-LASSO
age	0	0	0
smokstat	0	0	0
quetelet	0	-0.057	0
vituse	0	0	0
calories	0	0	0
fat	0	0	0
fiber	0.114	0.077	0.058
alcohol	0	0	0
cholesterol	0	0	0
betadiet	0	0	0.075
MAPE	0.559	0.503	0.568

Bootstrap Results

Dataset	Method	No. of non-zeros	Model Error	
			Median	MAD
BHP	ESL	3.710(0.830)	0.381(0.021)	0.180(0.069)
	CQR	7.025(1.015)	0.286(0.016)	0.258(0.020)
	LAD	5.020(0.839)	0.277(0.017)	0.113(0.075)
PBC	ESL	0.305(0.462)	0.459(0.030)	0.180(0.054)
	CQR	2.915(1.026)	0.453(0.032)	0.299(0.050)
	LAD	2.570(1.020)	0.429(0.036)	0.176(0.161)

Remarks

- Proposed a robust variable selection procedure via a penalized regression with the exponential squared loss.
- Investigated the sampling properties and studied the robustness properties of the proposed estimators.
- Illustrated that our estimators possessed the highest finite sample breakdown point, and the influence functions are bounded with respect to outliers in either the response or the covariate domain.

Remarks

- How to select both γ_n and regularization parameters λ_{nj} in a data-driven way is a difficult problem, since selection of γ_n depends on the choice of λ_{nj} and an estimate of $\boldsymbol{\beta}$.
- Although the data-adaptive methods such as cross-validation can be applied, it could cause huge computation and may not satisfy condition (C2).
- We chose regularization parameters λ_{nj} via a simple BIC criterion, and then proposed a data-driven approach to selecting the tuning parameter γ_n.
- We demonstrated the advantages of our methodology via simulation study and application.
- Our simulation studies revealed that the performance of ESL-LASSO is comparable to the oracle procedure irrespective of the presence and the mechanisms of outliers.

Acknowledgments

Heping Zhang's work was supported in part by grant R01 DA016750-09.

Thank you!

References

D.A. Belsley, E. Kuh, and R.E. Welsch. *Regression Diagnostics: Identifying Influential Data And Sources Of Collinearity*. Wiley, 1980.

J. Bradic, J. Fan, and W. Wang. Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):325–349, 2011.

J. Chen and Z. Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3): 759–771, 2008.

D.L. Donoho. Breakdown properties of multivariate location estimators. Technical report, Technical report, Harvard University, Boston., 1982.

D.L. Donoho and P.J. Huber. The notion of breakdown point. *A Festschrift for Erich L. Lehmann*, pages 157–184, 1983.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

IE Frank and JH Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 28 (2):337–407, 2000.

D. Gervini and V.J. Yohai. A class of robust and fully efficient regression estimators. *The Annals of Statistics*, 30(2):583–616, 2002.

F.R. Hampel. *Contributions to the theory of robust estimation*. PhD thesis, University of California Berkeley, 1968.

F.R. Hampel. A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, 42(6):1887–1896, 1971.

D. Harrison and D.L. Rubinfeld. Hedonic prices and the demand for clean air. *J. Environ. Economics and Management*, 5:81–102, 1978.

B.A. Johnson and L. Peng. Rank-based variable selection. *Journal of Nonparametric Statistics*, 20(3):241–252, 2008.

B. Kai, R. Li, and H. Zou. New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *The Annals of Statistics*, 39(1):305–332, 2011.

C. Leng. Variable selection and coefficient estimation via regularized rank regression. *Statistica Sinica*, 20:167–181, 2010.

P.J. Rousseeuw and V.J. Yohai. Robust regression by means of s-estimators. *Robust and Nonlinear Time Series*, 26:256–272, 1984.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:267–288, 1996.

P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1):387–423, 2009.

H. Wang, G. Li, and G. Jiang. Robust regression shrinkage and consistent variable selection through the lad-lasso. *Journal of Business and Economic Statistics*, 25(3):347–355, 2007.

L. Wang and R. Li. Weighted wilcoxon-type smoothly clipped absolute deviation method. *Biometrics*, 65(2):564–571, 2009.