Sparse Robust Regression using Non-concave Penalized Density Power Divergence

Subhabrata Majumdar
Joint work with Abhik Ghosh
University of Florida Informatics Institute

IISA-2018 conference, Gainesville, FL
May 19, 2018

Table of contents

Outline

Standard **linear regression model** (LRM):

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ are responses, $\mathbf{X} = (\boldsymbol{x}_1 \cdots \boldsymbol{x}_n)^T$ is the design matrix, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ are the random error components.

Penalized linear regression

Standard **linear regression model** (LRM):

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ are responses, $\mathbf{X} = (\boldsymbol{x}_1 \cdots \boldsymbol{x}_n)^T$ is the design matrix, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ are the random error components.

Sparse estimators of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, are defined as the minimizer of:

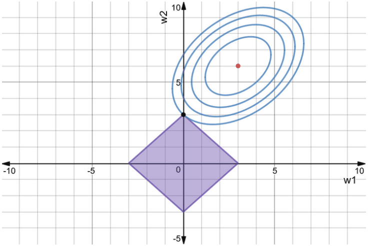
$$\sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^T \boldsymbol{\beta}) + \lambda_n \sum_{j=1}^{p} p(|\beta_j|),$$

where $\rho(.)$ is a loss function, $p(.)$ is the sparsity inducing penalty function, and $\lambda_n \equiv \lambda$ is the regularization parameter depending on n.

Sparse penalized least squares

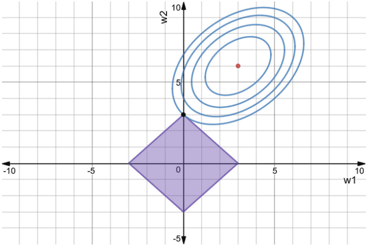
Linear model: $\mathbf{y} = \mathbf{X}\beta + \epsilon$, with $\mathbf{X} \in \mathbb{R}^{n \times p}, \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ with $\sigma > 0$;

Lasso (Tibshirani, 1996)
$\widehat{\beta} = \frac{1}{n} \operatorname{argmin}_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \|\boldsymbol{\beta}\|_1$;



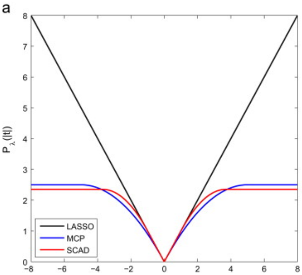
Linear model: $\mathbf{y} = \mathbf{X}\beta + \epsilon$, with $\mathbf{X} \in \mathbb{R}^{n \times p}, \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ with $\sigma > 0$;

Lasso (Tibshirani, 1996)
$\widehat{\beta} = \frac{1}{n} \operatorname{argmin}_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \|\boldsymbol{\beta}\|_1$;



SCAD (Fan and Li, 2001)
$\widehat{\beta} = \operatorname{argmin}_{\beta} \frac{1}{n} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \sum_{j=1}^{p} p(|\beta_j|)$;

MCP (Zhang, 2010)



- Sparse versions of robust regression methods- RLARS (Khan et al., 2007), sparse least trimmed squares (Wang et al., 2007), LAD-lasso (Alfons et al., 2013);
- Robust high-dimensional M-estimation- Neghaban et al. (2012); Bean et al. (2013); Donoho and Montanari (2016); Lozano et al. (2016); Loh and Wainwright (2017)

Why do we need another?

Why do we need another?

1. All methods until now focus on ℓ_1-penalization. But the bias of lasso-type estimators is well-known.

Why do we need another?

1. All methods until now focus on ℓ_1-penalization. But the bias of lasso-type estimators is well-known.

2. Many proposed methods lack theoretical rigor and only give algorithms.

Why do we need another?

1. All methods until now focus on ℓ_1-penalization. But the bias of lasso-type estimators is well-known.
2. Many proposed methods lack theoretical rigor and only give algorithms.
3. Robustness is either shown empirically or theoretically- not both.

Why do we need another?

1. All methods until now focus on ℓ_1-penalization. But the bias of lasso-type estimators is well-known.
2. Many proposed methods lack theoretical rigor and only give algorithms.
3. Robustness is either shown empirically or theoretically- not both.
4. Conditions assumed on the design matrix are largely similar to non-robust cases.

Example

$\mathbf{X}^T\mathbf{X}/n \to \mathbf{C}$ (Alfons et al., 2013)
Restricted eigenvalue condition (Lozano et al., 2016)

Outline

The DPD loss function

- **D**ensity **P**ower **D**ivergence is a generalization of the KL-divergence.

- DPD-based regression (Durio and Isaia, 2011) maximizes the loss function

$$L_n^\alpha(\boldsymbol{\beta}, \sigma) = \frac{1}{(2\pi)^{\alpha/2}\sigma^\alpha\sqrt{1+\alpha}} \left[1 - \frac{(1+\alpha)^{3/2}}{\alpha} \frac{1}{n} \sum_{i=1}^{n} e^{-\alpha \frac{(y_i - x_i^T \beta)^2}{2\sigma^2}} \right]$$

- **D**ensity **P**ower **D**ivergence is a generalization of the KL-divergence.
- DPD-based regression (Durio and Isaia, 2011) maximizes the loss function

$$L_n^\alpha(\boldsymbol{\beta}, \sigma) = \frac{1}{(2\pi)^{\alpha/2}\sigma^\alpha\sqrt{1+\alpha}}\left[1 - \frac{(1+\alpha)^{3/2}}{\alpha}\frac{1}{n}\sum_{i=1}^{n} e^{-\alpha\frac{(y_i - x_i^T\beta)^2}{2\sigma^2}}\right]$$

Why use DPD?

- **D**ensity **P**ower **D**ivergence is a generalization of the KL-divergence.
- DPD-based regression (Durio and Isaia, 2011) maximizes the loss function

$$L_n^\alpha(\boldsymbol{\beta}, \sigma) = \frac{1}{(2\pi)^{\alpha/2}\sigma^\alpha\sqrt{1+\alpha}} \left[1 - \frac{(1+\alpha)^{3/2}}{\alpha} \frac{1}{n} \sum_{i=1}^n e^{-\alpha \frac{(y_i - x_i^T \beta)^2}{2\sigma^2}} \right]$$

Why use DPD?

Adaptive: Large α = more robust, less efficient. Small α = more robust, less efficient.

- **D**ensity **P**ower **D**ivergence is a generalization of the KL-divergence.
- DPD-based regression (Durio and Isaia, 2011) maximizes the loss function

$$L_n^\alpha(\beta, \sigma) = \frac{1}{(2\pi)^{\alpha/2}\sigma^\alpha\sqrt{1+\alpha}}\left[1 - \frac{(1+\alpha)^{3/2}}{\alpha}\frac{1}{n}\sum_{i=1}^{n} e^{-\alpha\frac{(y_i - x_i^T\beta)^2}{2\sigma^2}}\right]$$

Why use DPD?

Adaptive: Large α = more robust, less efficient. Small α = more robust, less efficient.

Generalized: As $\alpha \downarrow 0$, $L_n^\alpha(\beta, \sigma)$ coincides (in a limiting sense) with the negative log-likelihood.
(why? think L-Hospital's rule.)

Penalized DPD

$$L_n^\alpha(\boldsymbol{\beta}, \sigma) + \sum_{j=1}^{p} p_\lambda(|\beta_j|)$$

where $p_\lambda(\cdot)$ is a penalty function (lasso, SCAD, MCP, ...).

Penalized DPD

$$L_n^\alpha(\boldsymbol{\beta}, \sigma) + \sum_{j=1}^{p} p_\lambda(|\beta_j|)$$

where $p_\lambda(\cdot)$ is a penalty function (lasso, SCAD, MCP, ...).

As $\alpha \downarrow 0$, this becomes the (non-robust) non-concave penalized negative log-likelihood.

Computational algorithm

Starting from $\hat{\boldsymbol{\beta}}, \hat{\sigma}$, Iteratively minimize the following:

$$
\begin{aligned}
R_\lambda^\alpha(\boldsymbol{\beta}) &= L_n^\alpha(\boldsymbol{\beta}, \hat{\sigma}) + \sum_{j=1}^{p} p_\lambda(|\beta_j|), \\
S^\alpha(\sigma) &= L_n^\alpha(\hat{\boldsymbol{\beta}}, \sigma).
\end{aligned}
$$

Computational algorithm

Starting from $\hat{\boldsymbol{\beta}}, \hat{\sigma}$, Iteratively minimize the following:

$$R_\lambda^\alpha(\boldsymbol{\beta}) = L_n^\alpha(\boldsymbol{\beta}, \hat{\sigma}) + \sum_{j=1}^{p} p_\lambda(|\beta_j|),$$

$$S^\alpha(\sigma) = L_n^\alpha(\hat{\boldsymbol{\beta}}, \sigma).$$

- Update $\boldsymbol{\beta}$ using a Concave-Convex Procedure (CCCP):

$$p_\lambda(|\beta_j|) = \tilde{J}_\lambda(|\beta_j|) + \lambda|\beta_j| \simeq \nabla\tilde{J}_\lambda(|\beta_j^c|)\beta_j + \lambda|\beta_j|$$

  where $\tilde{J}(\cdot)$ is differentiable and concave, $\boldsymbol{\beta}^c$ is a current solution.

- Update σ using gradient descent.

Updating $\hat{\beta}$ and $\hat{\sigma}$

Updating $\hat{\beta}$ and $\hat{\sigma}$

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \operatorname*{argmin}_{\boldsymbol{\beta}} \left\{ L_n^{\alpha}\left(\boldsymbol{\beta}, \hat{\sigma}^{(k)}\right) + \sum_{j=1}^{p} \left[\nabla \tilde{J}_{\lambda}(|\hat{\beta}_j^{(k)}|)\beta_j + \lambda|\beta_j| \right] \right\};$$

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \underset{\boldsymbol{\beta}}{\text{argmin}} \left\{ L_n^{\alpha}\left(\boldsymbol{\beta}, \hat{\sigma}^{(k)}\right) + \sum_{j=1}^{p} \left[\nabla \tilde{J}_{\lambda}(|\hat{\beta}_j^{(k)}|)\beta_j + \lambda|\beta_j| \right] \right\};$$

$$\hat{\sigma}^{2(k+1)} = \left[\sum_{i=1}^{n} w_i^{(k)} - \frac{\alpha}{(1+\alpha)^{3/2}} \right] \left[\sum_{i=1}^{n} w_i^{(k)} \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta}^{(k+1)} \right)^2 \right]^{-1},$$

$$w_i^{(k)} := \exp\left\{ -\alpha \frac{(y_i - \mathbf{x}_i^T \boldsymbol{\beta}^{(k)})^2}{\sigma^{2(k)}} \right\}.$$

Tuning parameter selection

To choose λ, we use a robust High-dimensional BIC:

$$\text{HBIC}(\lambda) = \log(\hat{\sigma}^2) + \frac{\log\log(n)\log p}{n}\|\hat{\beta}\|_0, \tag{1}$$

and select the optimal λ^* that minimizes the HBIC over a pre-determined set of values Λ_n: $\lambda^* = \text{argmin}_{\lambda\in\Lambda_n}\text{HBIC}(\lambda)$.

Outline

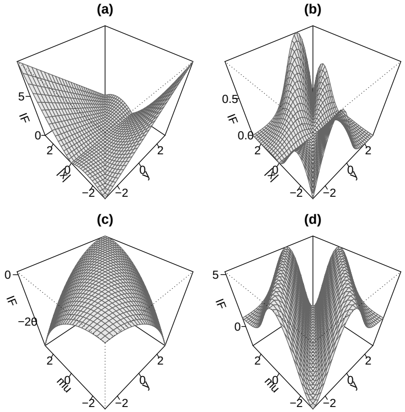
The **Influence Function (IF)** is a classical tool of measuring the asymptotic local robustness of any estimator (Hampel, 1968, 1974).

The **Influence Function (IF)** is a classical tool of measuring the asymptotic local robustness of any estimator (Hampel, 1968, 1974).

Consider a contaminated version of the true distribution joint G given by $G_\epsilon = (1-\epsilon)G + \epsilon\wedge_{(y_t,\boldsymbol{x}_t)}$ where ϵ is the contamination proportion and $\wedge_{(y_t,\boldsymbol{x}_t)}$ is the degenerate distribution at $(y_t, \boldsymbol{x}_t)$. Then, the IF of any functional $\boldsymbol{T}$ at G is defined as the *limiting (standardized) bias due to infinitesimal contamination*:

$$\mathcal{IF}((y_t, \boldsymbol{x}_t), \boldsymbol{T}_\alpha, G) = \lim_{\epsilon\to 0} \frac{\boldsymbol{T}_\alpha(G_\epsilon) - \boldsymbol{T}_\alpha(G)}{\epsilon} = \frac{\partial}{\partial\epsilon}\boldsymbol{T}_\alpha(G_\epsilon)\Big|_{\epsilon=0}.$$

IF for our estimates



Influence function plots for $\boldsymbol{\beta}$ (panels a and b, $(y_t, \|\mathbf{x}_{1t}\|_1)$ on the (x, y) axes, and ℓ_2 norms of IFs are plotted) and σ (panels c and d, $(y_t, \mathbf{x}_t^T\beta)$ on the axes). We assume $\mathbf{x}_{1t}$ is drawn from $\mathcal{N}_5(\mathbf{0}, \mathbf{I})$, and $\boldsymbol{\beta}_1 = (1, 1, 1, 1, 1)^T, \sigma = 1$. Panels a and c are for $\alpha = 0$, while b and d are for $\alpha = 0.5$

Outline

Modified conditions for robustness: example

Denote the non-zero index set of the true coefficient vector $\boldsymbol{\beta}^*$ by S.

- **Restricted eigenvalue condition**

$$\frac{\|\mathbf{X}\boldsymbol{\delta}\|^2}{n\|\boldsymbol{\delta}\|^2} \geq \kappa$$

for some $\kappa > 0$ and $\boldsymbol{\delta} \in \mathbb{R}^p$ s.t. $\|\boldsymbol{\delta}_{S^c}\|_1 \leq 3\|\boldsymbol{\delta}_S\|_1$.

Modified conditions for robustness: example

Denote the non-zero index set of the true coefficient vector $\boldsymbol{\beta}^*$ by S.

- **Restricted eigenvalue condition**

$$\frac{\|\mathbf{X}\boldsymbol{\delta}\|^2}{n\|\boldsymbol{\delta}\|^2} \geq \kappa$$

  for some $\kappa > 0$ and $\boldsymbol{\delta} \in \mathbb{R}^p$ s.t. $\|\boldsymbol{\delta}_{S^c}\|_1 \leq 3\|\boldsymbol{\delta}_S\|_1$.

- **Our condition**

$$\min_{(\boldsymbol{\delta},\sigma)\in\mathcal{N}_0} \Lambda_{\min}\left[\frac{1}{n}\mathbf{X}_S^T \nabla^2 L_n^\alpha(\boldsymbol{\delta},\sigma)\mathbf{X}_S\right] \geq c$$

  for $c > 0$, and

$$\mathcal{N}_0 = \left\{(\boldsymbol{\delta},\sigma) : \boldsymbol{\delta}_{S^c} = \mathbf{0}, \|(\boldsymbol{\delta}_S,\sigma) - (\boldsymbol{\beta}_S^*,\sigma^*)\|_\infty < \frac{\min_j|\beta_j^*|}{2}\right\}$$

Results

- Under a few conditions we prove that $\hat{\boldsymbol{\beta}}_{S^c} = \mathbf{0}$ and

$$\|\hat{\boldsymbol{\beta}}_S - \hat{\boldsymbol{\beta}}^*_S\|_\infty = O\left(\frac{\log n}{n^\tau}\right); \quad |\hat{\sigma} - \sigma^*| = O\left(\frac{\log n}{n^\tau}\right)$$

for some $0 < \tau < 0.5$.

- Under a few conditions we prove that $\hat{\beta}_{S^c} = \mathbf{0}$ and

$$\|\hat{\beta}_S - \hat{\beta}_S^*\|_\infty = O\left(\frac{\log n}{n^\tau}\right); \quad |\hat{\sigma} - \sigma^*| = O\left(\frac{\log n}{n^\tau}\right)$$

  for some $0 < \tau < 0.5$.

- These rates improve to $O(\sqrt{s/n})$ and $O(n^{-1/2})$ respectively under stronger conditions.

Results

- Under a few conditions we prove that $\hat{\boldsymbol{\beta}}_{S^c} = \mathbf{0}$ and

$$\|\hat{\boldsymbol{\beta}}_S - \hat{\boldsymbol{\beta}}_S^*\|_\infty = O\left(\frac{\log n}{n^\tau}\right); \quad |\hat{\sigma} - \sigma^*| = O\left(\frac{\log n}{n^\tau}\right)$$

  for some $0 < \tau < 0.5$.

- These rates improve to $O(\sqrt{s/n})$ and $O(n^{-1/2})$ respectively under stronger conditions.

- Under yet stronger conditions, we prove asymptotic normality.

Outline

Setup

Setup

- Obtain rows of $\mathbf{X}$ as $n = 100$ random draws from $\mathcal{N}(0, \Sigma_X)$, where Σ_X is a positive definite with $(i,j)^{\text{th}}$ element given by $0.5^{|i-j|}$.

Setup

- Obtain rows of $\mathbf{X}$ as $n = 100$ random draws from $\mathcal{N}(0, \Sigma_X)$, where Σ_X is a positive definite with $(i,j)^{\text{th}}$ element given by $0.5^{|i-j|}$.
- Given p, we consider two settings for $\boldsymbol{\beta}$:
  - Setting A (strong signal): For $j \in \{1, 2, 4, 7, 11\}$, $\beta_j = j$, otherwise 0;
  - Setting B (weak signal): Set $\beta_1 = \beta_7 = 1.5$, $\beta_2 = 0.5$, $\beta_4 = \beta_{11} = 1$, and 0 otherwise.

Setup

- Obtain rows of $\mathbf{X}$ as $n = 100$ random draws from $\mathcal{N}(0, \Sigma_X)$, where Σ_X is a positive definite with $(i, j)^{\text{th}}$ element given by $0.5^{|i-j|}$.
- Given p, we consider two settings for $\boldsymbol{\beta}$:
  - Setting A (strong signal): For $j \in \{1, 2, 4, 7, 11\}, \beta_j = j$, otherwise 0;
  - Setting B (weak signal): Set $\beta_1 = \beta_7 = 1.5, \beta_2 = 0.5, \beta_4 = \beta_{11} = 1$, and 0 otherwise.
- Generate the random errors as $\epsilon \sim N(0, 0.5^2)$, and set $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$.

Setup

- Obtain rows of $\mathbf{X}$ as $n = 100$ random draws from $\mathcal{N}(0, \Sigma_X)$, where Σ_X is a positive definite with $(i, j)^{\text{th}}$ element given by $0.5^{|i-j|}$.
- Given p, we consider two settings for $\boldsymbol{\beta}$:
  - Setting A (strong signal): For $j \in \{1, 2, 4, 7, 11\}, \beta_j = j$, otherwise 0;
  - Setting B (weak signal): Set $\beta_1 = \beta_7 = 1.5, \beta_2 = 0.5, \beta_4 = \beta_{11} = 1$, and 0 otherwise.
- Generate the random errors as $\epsilon \sim N(0, 0.5^2)$, and set $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.
- Three outlier settings:
  - Y-outliers: We add 20 to the response variables of a random 10% of samples,
  - X-outliers: We add 20 to each of the elements in the first 10 rows of $\mathbf{X}$ for a random 10% of samples,
  - No outliers.

Setup

- Obtain rows of $\mathbf{X}$ as $n = 100$ random draws from $\mathcal{N}(0, \Sigma_X)$, where Σ_X is a positive definite with $(i,j)^{\text{th}}$ element given by $0.5^{|i-j|}$.
- Given p, we consider two settings for $\boldsymbol{\beta}$:
    - Setting A (strong signal): For $j \in \{1, 2, 4, 7, 11\}$, $\beta_j = j$, otherwise 0;
    - Setting B (weak signal): Set $\beta_1 = \beta_7 = 1.5$, $\beta_2 = 0.5$, $\beta_4 = \beta_{11} = 1$, and 0 otherwise.
- Generate the random errors as $\epsilon \sim N(0, 0.5^2)$, and set $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$.
- Three outlier settings:
    - Y-outliers: We add 20 to the response variables of a random 10% of samples,
    - X-outliers: We add 20 to each of the elements in the first 10 rows of $\mathbf{X}$ for a random 10% of samples,
    - No outliers.

- Methods compared- RLARS, sLTS, RANSAC, LAD-Lasso, DPD-lasso, log DPD-lasso, Lasso, SCAD, MCP. We repeat model fitting by our method (DPD-ncv), DPD-lasso and LDPD-lasso for $\alpha = 0.2, 0.4, 0.6, 0.8, 1$, as well as for different values of the starting point, chosen by RLARS, sLTS and RANSAC.
- RLARS solution is used as our starting point.

Metrics

$$
\begin{aligned}
\mathrm{MSEE}(\hat{\boldsymbol{\beta}}) &= (1/p)\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2, \\
\mathrm{RMSPE}(\hat{\boldsymbol{\beta}}) &= \sqrt{\|\mathbf{y}_{test} - \mathbf{X}_{test}\hat{\boldsymbol{\beta}}\|^2}, \\
\mathrm{EE}(\widehat{\sigma}) &= |\hat{\sigma} - \sigma_0|, \\
\mathrm{TP}(\hat{\boldsymbol{\beta}}) &= \frac{|\operatorname{supp}(\widehat{\boldsymbol{\beta}}) \cap \operatorname{supp}(\boldsymbol{\beta}_0)|}{|\operatorname{supp}(\boldsymbol{\beta}_0)|}, \\
\mathrm{TN}(\hat{\boldsymbol{\beta}}) &= \frac{|\operatorname{supp}(\widehat{\boldsymbol{\beta}}) \cap \operatorname{supp}(\boldsymbol{\beta}_0)|}{|\operatorname{supp}(\boldsymbol{\beta}_0)|}, \\
\mathrm{MS}(\hat{\boldsymbol{\beta}}) &= |\operatorname{supp}(\widehat{\boldsymbol{\beta}})|.
\end{aligned}
$$

Table of outputs for $p = 500$ and Y-outliers

Setting B						
Method	MSEE($\hat{\beta}$) ($\times 10^{-4}$)	RMSPE($\hat{\beta}$) ($\times 10^{-2}$)	EE($\widehat{\sigma}$)	TP($\hat{\beta}$)	TN($\hat{\beta}$)	MS($\hat{\beta}$)
RLARS	1.1	4.58	0.09	1.00	1.00	6.00
sLTS	6.2	6.06	0.23	1.00	0.93	40.07
RANSAC	6.2	4.82	0.24	1.00	0.92	44.00
LAD-Lasso	68.6	15.65	2.77	0.65	0.99	6.28
DPD-ncv, $\alpha = 0.2$	0.8	4.28	0.06	1.00	1.00	5.00
DPD-ncv, $\alpha = 0.4$	0.8	4.30	0.06	1.00	1.00	5.00
DPD-ncv, $\alpha = 0.6$	0.8	4.50	0.06	1.00	1.00	5.00
DPD-ncv, $\alpha = 0.8$	0.7	4.59	0.06	1.00	1.00	5.00
DPD-ncv, $\alpha = 1$	0.8	4.61	0.06	1.00	1.00	5.00
DPD-Lasso, $\alpha = 0.2$	61.3	15.10	0.05	1.00	0.00	499.08
DPD-Lasso, $\alpha = 0.4$	58.9	14.41	0.17	1.00	0.05	477.15
DPD-Lasso, $\alpha = 0.6$	56.5	14.85	0.14	1.00	0.10	450.22
DPD-Lasso, $\alpha = 0.8$	55.1	14.29	0.02	1.00	0.13	435.72
DPD-Lasso, $\alpha = 1$	54.2	14.16	0.01	1.00	0.13	433.65
LDPD-Lasso, $\alpha = 0.2$	2.1	5.09	0.07	1.00	0.99	10.19
LDPD-Lasso, $\alpha = 0.4$	2.2	5.12	0.09	1.00	0.99	7.97
LDPD-Lasso, $\alpha = 0.6$	2.3	5.14	0.11	1.00	0.99	7.62
LDPD-Lasso, $\alpha = 0.8$	2.3	5.14	0.13	1.00	1.00	7.38
LDPD-Lasso, $\alpha = 1$	2.3	5.15	0.14	1.00	1.00	7.38
Lasso	134.1	22.41	4.54	0.02	1.00	0.24
SCAD	128.6	20.97	3.60	0.32	0.99	8.72
MCP	141.6	21.09	3.69	0.24	0.99	4.52

Table of outputs for $p = 500$ and X-outliers

Setting B						
Method	MSEE($\hat{\beta}$) ($\times 10^{-4}$)	RMSPE($\hat{\beta}$) ($\times 10^{-2}$)	EE($\widehat{\sigma}$)	TP($\hat{\beta}$)	TN($\hat{\beta}$)	MS($\hat{\beta}$)
RLARS	2.0	4.2	0.14	1.00	0.99	12.00
sLTS	8.7	5.3	0.24	1.00	0.92	42.50
RANSAC	5.8	5.9	0.26	1.00	0.98	15.00
LAD-Lasso	108.0	20.4	2.87	0.38	0.99	7.71
DPD-ncv, $\alpha = 0.2$	1.2	4.1	0.08	1.00	1.00	7.00
DPD-ncv, $\alpha = 0.4$	1.1	4.0	0.10	1.00	1.00	7.00
DPD-ncv, $\alpha = 0.6$	1.1	4.2	0.12	1.00	1.00	7.00
DPD-ncv, $\alpha = 0.8$	1.4	4.2	0.14	1.00	1.00	7.00
DPD-ncv, $\alpha = 1$	1.5	4.2	0.15	1.00	1.00	7.00
DPD-Lasso, $\alpha = 0.2$	59.5	13.8	0.05	1.00	0.01	495.26
DPD-Lasso, $\alpha = 0.4$	48.6	10.8	0.20	1.00	0.16	420.56
DPD-Lasso, $\alpha = 0.6$	35.3	9.2	0.28	1.00	0.35	329.12
DPD-Lasso, $\alpha = 0.8$	27.6	8.6	0.13	1.00	0.45	278.17
DPD-Lasso, $\alpha = 1$	25.7	9.2	0.01	1.00	0.47	267.29
LDPD-Lasso, $\alpha = 0.2$	1.9	5.0	0.06	1.00	0.98	15.14
LDPD-Lasso, $\alpha = 0.4$	1.8	5.0	0.07	1.00	0.98	14.04
LDPD-Lasso, $\alpha = 0.6$	1.8	5.1	0.07	1.00	0.98	14.03
LDPD-Lasso, $\alpha = 0.8$	1.8	5.0	0.07	1.00	0.98	14.47
LDPD-Lasso, $\alpha = 1$	1.8	5.0	0.07	1.00	0.98	13.90
LASSO	22.6	10.3	0.13	0.99	0.87	70.32
SCAD	45.8	13.8	0.55	0.81	0.98	16.25
MCP	45.2	12.8	0.49	0.81	0.97	16.45

Table of outputs for $p = 500$ and no outliers

Setting B						
Method	MSEE($\hat{\beta}$) ($\times 10^{-4}$)	RMSPE($\hat{\beta}$) ($\times 10^{-2}$)	EE($\widehat{\sigma}$)	TP($\hat{\beta}$)	TN($\hat{\beta}$)	MS($\hat{\beta}$)
RLARS	1.4	4.73	0.12	1.00	0.99	10.00
sLTS	7.9	5.65	0.24	1.00	0.93	42.00
RANSAC	5.2	4.95	0.23	1.00	0.98	15.00
LAD-Lasso	4.7	3.90	0.42	1.00	1.00	7.30
DPD-ncv, $\alpha = 0.2$	1.4	4.73	0.12	1.00	0.99	10.00
DPD-ncv, $\alpha = 0.4$	1.4	4.73	0.12	1.00	0.99	10.00
DPD-ncv, $\alpha = 0.6$	1.4	4.73	0.12	1.00	0.99	10.00
DPD-ncv, $\alpha = 0.8$	1.4	4.73	0.12	1.00	0.99	10.00
DPD-ncv, $\alpha = 1$	1.4	4.73	0.12	1.00	0.99	10.00
DPD-Lasso, $\alpha = 0.2$	79.1	14.56	0.10	1.00	0.00	499.00
DPD-Lasso, $\alpha = 0.4$	58.2	12.98	0.25	1.00	0.14	429.70
DPD-Lasso, $\alpha = 0.6$	44.9	10.18	0.25	1.00	0.31	348.80
DPD-Lasso, $\alpha = 0.8$	19.9	8.86	0.05	1.00	0.57	215.60
DPD-Lasso, $\alpha = 1$	21.1	11.46	0.00	1.00	0.59	208.70
LDPD-Lasso, $\alpha = 0.2$	1.9	3.94	0.06	1.00	0.97	17.50
LDPD-Lasso, $\alpha = 0.4$	2.0	4.05	0.09	1.00	0.98	15.50
LDPD-Lasso, $\alpha = 0.6$	2.0	4.23	0.09	1.00	0.98	16.90
LDPD-Lasso, $\alpha = 0.8$	2.0	4.16	0.08	1.00	0.98	16.00
LDPD-Lasso, $\alpha = 1$	2.0	4.10	0.08	1.00	0.98	16.40
Lasso	2.1	3.59	0.33	1.00	0.98	12.90
SCAD	0.3	3.71	0.21	1.00	0.99	9.70
MCP	0.3	3.69	0.20	1.00	1.00	6.80

Conclusion

- We proposed a sparse regression method based on a generalization of the log-likelihood;
- We provide detailed theoretical analysis for the robustness and consistency properties of estimates of β and σ;
- Future directions- robust high-dimensional testing for β, graphical models, group sparsity.

Conclusion

- We proposed a sparse regression method based on a generalization of the log-likelihood;
- We provide detailed theoretical analysis for the robustness and consistency properties of estimates of β and σ;
- Future directions- robust high-dimensional testing for β, graphical models, group sparsity.

Conclusion

- We proposed a sparse regression method based on a generalization of the log-likelihood;
- We provide detailed theoretical analysis for the robustness and consistency properties of estimates of $\boldsymbol{\beta}$ and σ;
- Future directions- robust high-dimensional testing for β, graphical models, group sparsity.

Conclusion

- We proposed a sparse regression method based on a generalization of the log-likelihood;
- We provide detailed theoretical analysis for the robustness and consistency properties of estimates of $\boldsymbol{\beta}$ and σ;
- Future directions- robust high-dimensional testing for $\boldsymbol{\beta}$, graphical models, group sparsity.

References

Preprint available at: https://arxiv.org/abs/1803.03348

Alfons, A., Croux, C., and Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Ann. Appl. Statist.*, 7:226–248.

Bean, D., Bickel, P., El Karoui, N., and Yu, B. (2013). Optimal M-estimation in high-dimensional regression. *Proc. Natl. Acad. Sci.*, 110(36):14563–14568.

Donoho, D. and Montanari, A. (2016). High dimensional robust M-estimation: asymptotic variance via approximate message passing. *Probab. Theory Relat. Fields*, 166:935–969.

Durio, A. and Isaia, E. D. (2011). The minimum density power divergence approach in building robust regression models. *Informatica*, 22(1):43–56.

Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *J. Amer. Statist. Assoc.*, 96:1348–1360.

Hampel, F. R. (1968). *Contributions to the Theory of Robust Estimation*. Ph.d. thesis, University of California, Berkeley, USA.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist Assoc.*, 69:383–393.

Khan, J. A., van Aelst, S., and Zamar, R. H. (2007). Robust linear model selection based on least angle regression. *J. Amer. Statist. Assoc.*, 102:1289–1299.

Loh, P.-L. and Wainwright, M. J. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *Ann. Statist.*, 45(2):866–896.

Lozano, A., Meinshausen, N., and Yang, E. (2016). Minimum Distance Lasso for robust high-dimensional regression. *Electron. J. Stat.*, 10:1296–1340.

Neghaban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012). A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers. *Stat. Sci.*, 27(4):538–557.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, 58(267–288).

Wang, H., Li, G., and Jiang, G. (2007). Robust Regression Shrinkage and Consistent Variable Selection Through the LAD-Lasso. *J. Bus. Econ. Stat.*, 25(3):347–355.

Zhang, C. H. (2010). Nearly Unbiased Variable Selection under Minimax Concave Penalty. *Ann. Statist.*, 38:894–942.

Preprint available at: https://arxiv.org/abs/1803.03348

Alfons, A., Croux, C., and Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Ann. Appl. Statist.*, 7:226–248.

Bean, D., Bickel, P., El Karoui, N., and Yu, B. (2013). Optimal M-estimation in high-dimensional regression. *Proc. Natl. Acad. Sci.*, 110(36):14563–14568.

Donoho, D. and Montanari, A. (2016). High dimensional robust M-estimation: asymptotic variance via approximate message passing. *Probab. Theory Relat. Fields*, 166:935–969.

Durio, A. and Isaia, E. D. (2011). The minimum density power divergence approach in building robust regression models. *Informatica*, 22(1):43–56.

Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *J. Amer. Statist. Assoc.*, 96:1348–1360.

Hampel, F. R. (1968). *Contributions to the Theory of Robust Estimation*. Ph.d. thesis, University of California, Berkeley, USA.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist Assoc.*, 69:383–393.

Khan, J. A., van Aelst, S., and Zamar, R. H. (2007). Robust linear model selection based on least angle regression. *J. Amer. Statist. Assoc.*, 102:1289–1299.

Loh, P.-L. and Wainwright, M. J. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *Ann. Statist.*, 45(2):866–896.

Lozano, A., Meinshausen, N., and Yang, E. (2016). Minimum Distance Lasso for robust high-dimensional regression. *Electron. J. Stat.*, 10:1296–1340.

Neghaban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012). A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers. *Stat. Sci.*, 27(4):538–557.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, 58(267–288).

Wang, H., Li, G., and Jiang, G. (2007). Robust Regression Shrinkage and Consistent Variable Selection Through the LAD-Lasso. *J. Bus. Econ. Stat.*, 25(3):347–355.

Zhang, C. H. (2010). Nearly Unbiased Variable Selection under Minimax Concave Penalty. *Ann. Statist.*, 38:894–942.

THANK YOU!