Robust High-dimensional Statistics III

Adaptive Procedures & other Practical Considerations

Dr. Abhik Ghosh

Indian Statistical Institute, Kolkata, India.

2022

Table of contents

1. Robust Adaptive Procedures
2. Robust Variable Screening
3. Stability of the Set of Selected Variables
4. Conclusion

Outline

1. **Robust Adaptive Procedures**
2. Robust Variable Screening
3. Stability of the Set of Selected Variables
4. Conclusion

Adaptive Procedures to Reduce False Positives

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 $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ are the random error components.

Adaptive LASSO (Zou, 2006)

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^{p} \frac{|\beta_j|}{|\widetilde{\beta}_{j,init}|} \right\}$$

Advantages: **Variable Selection Consistency; Reduced bias & false discovery**.

- Zou (2006) proved its oracle property for the low-dimensional cases with fixed p.
- Huang et al. (2008) derived its oracle consistency for high-dimensional set-ups ($p \gg n$).

Main Result (Huang et al., 2008)

Under suitable assumptions under high-dimensional set-up $p \gg n$, if the initial estimator $\widetilde{\boldsymbol{\beta}}_{init}$ used is r_n-consistent, then the adaptive LASSO estimate $\widehat{\boldsymbol{\beta}}$ satisfies the following properties:

- $\mathbf{P}\left(\mathbf{sign}(\widehat{\boldsymbol{\beta}}) = \mathbf{sign}(\widehat{\boldsymbol{\beta}}_0)\right) \to 1$.
- If $r_n \lambda n^{-1/2} \to 0$, then for any $\boldsymbol{u} \in \mathbb{R}^s$ with $\boldsymbol{u}^T\boldsymbol{u} \leq 1$, $\boldsymbol{u}^{\mathbf{T}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\mathbf{0}}) \xrightarrow{\mathcal{D}} \mathbf{N}\left(\mathbf{0}, \sigma^2 \boldsymbol{u}^{\mathbf{T}}(\mathbf{X}_{\mathbf{S}}^{\mathbf{T}}\mathbf{X}_{\mathbf{S}})^{-1}\boldsymbol{u}\right)$.

Adaptive Penalty as An Approximation for Non-concave Penalties

A More General Adaptive LASSO

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^{p} w(|\widetilde{\beta}_{j,init}|)|\beta_j| \right\}, \qquad w \text{ is a weight function!}$$

Approximation for Non-cocave Penlaty

$$p_{\lambda_n}(|\beta_j|) \approx p_{\lambda_n}(|\widetilde{\beta}_j|) + p'_{\lambda_n}(|\widetilde{\beta}_j|)(|\beta_j| - |\widetilde{\beta}_j|).$$

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + \sum_{j=1}^{p} p_{\lambda_n}(|\beta_j|) \right\} \\
&\approx \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + \sum_{j=1}^{p} p'_{\lambda_n}(|\widetilde{\beta}_{j,init}|)|\beta_j| \right\}
\end{aligned}$$

It is the general Adaptive Lasso estimator with $w = \frac{1}{\lambda_n} p'_{\lambda_n}$.

For SCAD: $$w(s) = \frac{1}{\lambda_n} p'_{\lambda_n}(s) = I(s \le \lambda_n) + \frac{(a\lambda_n - s)_+}{(a-1)\lambda_n} I(s > \lambda_n), \quad a > 2.$$

M-estimators with Adaptive Lasso Penalty

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 $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ are the random error components.

Adaptively Penalized M-estimators

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^T \boldsymbol{\beta}) + \lambda_n \sum_{j=1}^{p} w(|\widetilde{\beta}_{j,init}|)|\beta_j| \right\}$$

- Fan et al. (2014): $\rho_\tau(r) = r\{\tau - I(r \leq 0)\}$. [*Quantile loss*]
- Lambert-Lacroix and Zwald (2011, 2016): $\rho_k(r) = \frac{r^2}{2} I(|r| \leq k) + k(|r| - k/2) I(|r| > k)$. [*Huber loss*]
- Zheng et al. (2017): $\rho(r) = \alpha r^2 + (1-\alpha)|r|$ with $\alpha \in [0,1]$! (*Adaptive LAD-LASSO*?)
- Smucler and Yohai (2017): ρ as the MM-regression loss function!
- Chang et al. (2018): $\rho_k(r) = \frac{k^2}{6} \left\{ 1 - \left(1 - \frac{r^2}{k^2} \right)^3 \right\} I(|r| \leq k) + \frac{k^2}{6} I(|r| \geq k)$. [*Tukey's biweight loss*]
- Avella-Media and Ronchetti (2018): ρ as the Quasi-likelihood for GLM!

Robust Adaptive LASSO with DPD-based Loss

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,

We assume the random error components are IID with $\epsilon_i \sim \frac{1}{\sigma} f\left(\frac{\epsilon}{\sigma}\right)$,
where f is any univariate density with mean 0 and variance 1, such that $M_f^{(\alpha)} = \int f(\epsilon)^{1+\alpha} d\epsilon < \infty$.

The DPD-based loss functon (Ghosh and Majumdar, 2020)

$$\mathbf{L_n^{(\alpha)}}(\boldsymbol{\beta}, \sigma) = \frac{\mathbf{1}}{\sigma^{\alpha}} \mathbf{M_f^{(\alpha)}} - \frac{\mathbf{1}+\alpha}{\alpha} \frac{\mathbf{1}}{\mathbf{n}\sigma^{\alpha}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{f}^{\alpha}\left(\frac{\mathbf{y_i} - \boldsymbol{x_i^t}\boldsymbol{\beta}}{\sigma}\right) + \frac{\mathbf{1}}{\alpha}.$$

Adaptively Weighted DPD-LASSO (Ghosh et al., 2020)

$$\text{AW-DPD-LASSO estimator} : \left(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}\right) = \arg\min_{(\boldsymbol{\beta},\sigma)} \left\{ L_n^{(\alpha)}(\boldsymbol{\beta}, \sigma) + \lambda_n \sum_{j=1}^{p} w(|\widetilde{\beta}_{j,init}|)|\beta_j| \right\}$$

- **DPD-LASSO**: $w(u) = 1$ for all u.
- **Ad-DPD-LASSO**: $w(u) = \frac{1}{|u|} I(u \neq 0)$.

Computational Algorithm

$$\left(\widehat{\beta}, \widehat{\sigma}\right) = \arg\min_{(\beta,\sigma)} \left\{ L_n^{(\alpha)}(\beta, \sigma) + \lambda_n \sum_{j=1}^{p} w(|\widetilde{\beta}_{j,init}|)|\beta_j| \right\}, \qquad \text{with normal errors.}$$

Main idea: Update β and σ sequentially!

Update $\widehat{\beta}$ using Majorization technique (MM-algorithm)

$$\widehat{\boldsymbol{\beta}}^{(k+1)} = \text{argmin} \left[\sum_{i=1}^{n} \mu_i^{(k)} \left(\frac{y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}}{\widehat{\sigma}^{(k)}} \right)^2 + \lambda \sum_{j=1}^{p} w(|\tilde{\beta}_j|)|\beta_j| \right] \tag{1}$$

with

$$\mu_i^{(k)} = \exp\left(-\frac{\alpha}{2} \left(\frac{y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}^{(k)}}{\widehat{\sigma}^{(k)}} \right)^2 \right) \left[\sum_{l=1}^{n} \exp\left(-\frac{\alpha}{2} \left(\frac{y_l - \boldsymbol{x}_l^T \widehat{\boldsymbol{\beta}}^{(k)}}{\widehat{\sigma}^{(k)}} \right)^2 \right) \right]^{-1}.$$

Update σ using gradient descent

$$\widehat{\sigma}^{2(k+1)} = \frac{\left[\sum_{i=1}^{n} \omega_i^{(k)} - \frac{\alpha}{(1+\alpha)^{3/2}} \right]}{\left[\sum_{i=1}^{n} \omega_i^{(k)} \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta}^{(k+1)} \right)^2 \right]}, \qquad \omega_i^{(k)} := \exp\left\{ -\alpha \left(\frac{y_l - \boldsymbol{x}_l^T \widehat{\boldsymbol{\beta}}^{(k)}}{\widehat{\sigma}^{(k)}} \right)^2 \right\}.$$

Selection of λ: Minimize robust High-dimensional BIC, $\text{RobHBIC}(\lambda) = \log(\widehat{\sigma}^2) + \frac{\log\log(n)\log p}{n} \|\widehat{\boldsymbol{\beta}}\|_0$.

Statistical Functional for AW-DPD-LASSO

AW-DPD-LASSO Functional

$$\boldsymbol{T}_\alpha(G) = \left(\boldsymbol{T}_\alpha^\beta, T_\alpha^\sigma\right) = \arg\min_{(\boldsymbol{\beta},\sigma)} \left\{ \int L_\alpha^*((y,\boldsymbol{x});\boldsymbol{\theta}) dG(y,\boldsymbol{x}) + \lambda_n \sum_{j=1}^{p} w(|\boldsymbol{U}(G)|)|\beta_j| \right\}$$

where

$$L_\alpha^*((y,\boldsymbol{x});\boldsymbol{\theta}) = \frac{1}{\sigma^\alpha} M_f^{(\alpha)} - \frac{1+\alpha}{\alpha}\frac{1}{\sigma^\alpha} f^\alpha\left(\frac{y-\boldsymbol{x}^T\boldsymbol{\beta}}{\sigma}\right) + \frac{1}{\alpha}.$$

$$\psi_\alpha((y,\boldsymbol{x});\boldsymbol{\theta}) = \nabla L_\alpha^*((y,\boldsymbol{x});\boldsymbol{\theta}) = \frac{(1+\alpha)}{\sigma^{\alpha+1}} \begin{bmatrix} \psi_{1,\alpha}\left(\frac{y-\boldsymbol{x}^T\boldsymbol{\beta}}{\sigma}\right)\boldsymbol{x} \\ \psi_{2,\alpha}\left(\frac{y-\boldsymbol{x}^T\boldsymbol{\beta}}{\sigma}\right) \end{bmatrix}, \tag{2}$$

where $M_f^{(\alpha)} = \int f^{1+\alpha}$ and

$$\begin{aligned} \psi_{1,\alpha}(s) &= u(s)f^\alpha(s), \\ \psi_{2,\alpha}(s) &= \{su(s)+1\}f^\alpha(s) - \frac{\alpha}{\alpha+1} M_f^{(\alpha)}. \end{aligned} \tag{3}$$

Robustness: Influence Function of AW-DPD-LASSO

Theorem

Let us denote $\boldsymbol{\theta}_g = (\boldsymbol{\beta}_{1g}^T, \mathbf{0}_{p-s}^T, \sigma_g)^T = \boldsymbol{T}_\alpha(G) = (\boldsymbol{T}_{1,\alpha}^\beta(G)^T, \boldsymbol{T}_{2,\alpha}^\beta(G)^T, T_\alpha^\sigma(G))^T$. *Then, whenever the associated quantities exists, the influence function of* $\boldsymbol{T}_{2,\alpha}^\beta$ *is identically zero at* G *and that of* $(\boldsymbol{T}_{1,\alpha}^\beta, T_\alpha^\sigma)$ *at* G *is given by*

$$\mathcal{IF}((y_t, \boldsymbol{x}_t), (\boldsymbol{T}_{1,\alpha}^\beta, T_\alpha^\sigma), G) = -\boldsymbol{S}_\alpha^{-1} \begin{bmatrix} \frac{(1+\alpha)}{(\sigma^g)^{\alpha+1}} \psi_{1,\alpha}\left(\frac{y_t - \boldsymbol{x}_{1t}^T \beta^g}{\sigma^g}\right) \boldsymbol{x}_{1t} + \lambda \boldsymbol{P}^*(\boldsymbol{\beta}, \boldsymbol{U}(G)) \\ \qquad + \lambda \boldsymbol{P}^{**}(\boldsymbol{\beta}, \boldsymbol{U}(G)) \mathcal{IF}((y_t, \boldsymbol{x}_t), \boldsymbol{U}_1, G) \\ \frac{(1+\alpha)}{(\sigma^g)^{\alpha+1}} \psi_{2,\alpha}\left(\frac{y_t - \boldsymbol{x}_{1t}^T \beta^g}{\sigma^g}\right) \end{bmatrix},$$

where

$$\boldsymbol{P}^*(\boldsymbol{\beta}, \boldsymbol{u}) = \left(w(|u_j(G)|)\,\mathrm{sign}(\beta_j)\right)_{j=1,\ldots,s}, \quad \boldsymbol{P}^{**}(\boldsymbol{\beta}, \boldsymbol{u}) = Diag\left\{w'(|u_j(G)|) sign(U_j(G)\beta_j) : j = 1, \ldots,\right.$$

and

$$\boldsymbol{S}_\alpha(G; \beta, \sigma) = -\frac{(1+\alpha)}{\sigma^{\alpha+2}} E_G \begin{bmatrix} J_{11,\alpha}\left(\frac{y - \boldsymbol{x}^T\beta}{\sigma}\right) \boldsymbol{x}_1 \boldsymbol{x}_1^T & J_{12,\alpha}\left(\frac{y - \boldsymbol{x}^T\beta}{\sigma}\right) \boldsymbol{x}_1 \\ J_{12,\alpha}\left(\frac{y - \boldsymbol{x}^T\beta}{\sigma}\right) \boldsymbol{x}_1^T & J_{22,\alpha}\left(\frac{y - \boldsymbol{x}^T\beta}{\sigma}\right) \end{bmatrix}. \tag{4}$$

Example: IF of Ad-DPD-LASSO with Normal Error

Let us denote $\boldsymbol{\theta}_g = (\boldsymbol{\beta}_{1g}^T, \mathbf{0}_{p-s}^T, \sigma_g)^T = \boldsymbol{T}_\alpha(G) = (\boldsymbol{T}_{1,\alpha}^\beta(G)^T, \boldsymbol{T}_{2,\alpha}^\beta(G)^T, T_\alpha^\sigma(G))^T$. Then, whenever the associated quantities exists, then we have

$$\mathcal{IF}((y_t, \boldsymbol{x}_t), \boldsymbol{T}_{1,\alpha}^\beta, F_{(\boldsymbol{\beta},\sigma)}) = -\sigma(\alpha+1)^{3/2}\left[E_H\left(\boldsymbol{x}_1\boldsymbol{x}_1^T\right)\right]^{-1}\left[\left(y_t - \boldsymbol{x}_t^T\boldsymbol{\beta}\right)e^{-\frac{\alpha\left(y_t-\boldsymbol{x}_t^T\boldsymbol{\beta}\right)^2}{2\sigma^2}}\boldsymbol{x}_{1t}\right.$$
$$\left. + \frac{\lambda\sigma^{2\alpha+3}(2\pi)^{\alpha/2}}{(1+\alpha)}\left(\boldsymbol{P}_0^*(\boldsymbol{\beta}) + \boldsymbol{P}_0^{**}(\boldsymbol{\beta})\mathcal{IF}((y_t, \boldsymbol{x}_t), \boldsymbol{U}_1, F_{(\boldsymbol{\beta},\sigma)})\right)\right],$$

$$\mathcal{IF}((y_t, \boldsymbol{x}_t), \boldsymbol{T}_{2,\alpha}^\beta, F_{(\boldsymbol{\beta},\sigma)}) = 0,$$

$$\mathcal{IF}((y_t, \boldsymbol{x}_t), T_\alpha^\sigma, F_{\boldsymbol{\theta}}) = \frac{\sigma(1+\alpha)^{5/2}}{2+\alpha^2}\left[\left(1 - \left(\frac{y_t - \boldsymbol{x}_t^T\boldsymbol{\beta}}{\sigma}\right)^2\right)e^{-\frac{\alpha\left(y_t-\boldsymbol{x}_t^T\boldsymbol{\beta}\right)^2}{2\sigma^2}} - \frac{\alpha}{(1+\alpha)^{1/2}}\right],$$

where

$$\boldsymbol{P}_0^*(\boldsymbol{\beta}) = \left(|\beta_1|^{-1}, \ldots, |\beta_s|^{-1}\right), \qquad \boldsymbol{P}_0^{**}(\boldsymbol{\beta}) = \text{diag}\{\beta_1^{-2}, \ldots, \beta_s^{-2}\}.$$

As $\alpha \downarrow 0$, Ad-DPD-LASSO coincides with the usual adaptive LASSO and hence the above results also provide its influence function (unbounded) – new in the literature of adaptive LASSO!

Oracle Consistency: Fixed Non-stochastic Weights

Assume that the design matrix $\mathbf{X}$ is fixed with each column being standardized to have ℓ_1-norm $\sqrt{n}$ and the response is also standardized so that $\sigma^2 = 1$. Define $\delta_n = \sqrt{s(\log n)/n} + \lambda_n ||\boldsymbol{w}_0||_2$.

Theorem (Ghosh et al., 2020)

If Assumptions (A1)–(A4) hold and $\lambda_n ||\boldsymbol{w}_0||_2 \sqrt{s}\kappa_n \to 0$, *then, given any constant* $C_1 > 0$, *there exists some* $c > 0$ *such that any oracle AW-DPD-LASSO estimator* $\widehat{\boldsymbol{\beta}} = ((\widehat{\boldsymbol{\beta}}_1^o)^T, \mathbf{0}^T)^T$ *satisfies*

$$P\left(\left|\left|\widehat{\boldsymbol{\beta}}_1^o - \boldsymbol{\beta}_{10}\right|\right|_2 \leq C_1 \delta_n\right) \geq 1 - n^{-cs}. \tag{5}$$

Further, if $\delta_n^{-1} \min_{1 \leq j \leq s} |\beta_{j0}| \to \infty$, *then* **sign($\widehat{\boldsymbol{\beta}}_1^o$) = sign($\boldsymbol{\beta}_{10}$)**.

Theorem (Ghosh et al., 2020)

Suppose that Assumptions (A1)–(A5) hold with $\lambda_n > 2\sqrt{(c+1)\log p/n}$, $\min(|\boldsymbol{w}_1|) > c_3$, *and*

$$\lambda_n ||\boldsymbol{w}_0||_2 \kappa_n \max\{\sqrt{s}, ||\boldsymbol{w}_0||_2\} \to 0, \quad \delta_n s^{3/2} \kappa_n^2 (\log_2 n)^2 = o(n\lambda_n^2).$$

Then, with probability at least $1 - O(n^{-cs})$, *there exists a global minimizer* $\widehat{\boldsymbol{\beta}} = \left((\widehat{\boldsymbol{\beta}}_1^o)^T, \widehat{\boldsymbol{\beta}}_2^T\right)^T$ *of the AW-DPD-LASSO objective function such that*

$$\left|\left|\widehat{\boldsymbol{\beta}}_1^o - \boldsymbol{\beta}_{10}\right|\right|_2 \leq C_1 \delta_n, \quad \textit{and} \quad \widehat{\boldsymbol{\beta}}_2 = \mathbf{0}_{p-s}.$$

List of Assumptions

Given Standarized variables with $\sigma^2 = 1$:

$$\text{AW-DPD-LASSO estimator } : \widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ L_n^{(\alpha)}(\boldsymbol{\beta}) + \lambda_n \sum_{j=1}^{p} w(|\widetilde{\beta}_{j,init}|)|\beta_j| \right\}$$

$$\text{where } L_n^{(\alpha)}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \rho_\alpha(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}), \quad \text{with } \rho_\alpha(r) = M_f^{(\alpha)} - \frac{1+\alpha}{\alpha} f^\alpha(r) + \frac{1}{\alpha}.$$

Define: $\boldsymbol{H}_\alpha^{(1)}(\boldsymbol{\beta}) = (\rho'_\alpha(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}) : i = 1, \ldots, n)^T$, $\boldsymbol{H}_\alpha^{(2)}(\boldsymbol{\beta}) = \text{Diag}\{\rho''_\alpha(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}) : i = 1, \ldots, n\}$.

(A1) The error density f is such that f^α is Lipschitz with the Lipschitz constant L_α.

(A2) The eigenvalues of $n^{-1}(\boldsymbol{X}_1^T\boldsymbol{X}_1)$ are bounded below and above by positive constants c_0 and c_0^{-1}, respectively. Also $\kappa_n := \max_{i,j}|x_{ij}| = o(n^{1/2}s^{-1})$.

(A3) The diagonal elements of $E[\boldsymbol{H}_\alpha^{(2)}(\boldsymbol{\beta}_0)]$ are all finite and bounded from below by $c_1 > 0$.

(A4) Expectation of third order partial derivatives of $\rho_\alpha(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})$, $i = 1, \ldots, n$, with respect to all components of $\boldsymbol{\beta}_{S_0}$ are uniformly bounded in a neighborhood of $\boldsymbol{\beta}_{10}$.

(A5) $\left|\left| n^{-1}(\boldsymbol{X}_2^T E[\boldsymbol{H}_\alpha^{(2)}(\boldsymbol{\beta}_0)]\boldsymbol{X}_1) \right|\right|_{2,\infty} < \frac{\lambda_n \min(|\boldsymbol{w}_1|)}{2C_1\delta_n}$ for $C_1 > 0$; $||\boldsymbol{A}||_{2,\infty} = \sup_{\boldsymbol{x}\in\mathbb{R}^q\setminus\{0\}} \frac{||\boldsymbol{A}\boldsymbol{x}||_\infty}{||\boldsymbol{x}||_2}$.

Oracle Consistency: Stochastic Weights

(A7) The initial estimator $\widetilde{\beta}$ satisfies $||\widetilde{\beta} - \beta_0||_2 \leq C_2\sqrt{s(\log p)/n}$, with probability $\to 1$.

(A8) The weight function $w(\cdot)$ is non-increasing over $(0, \infty)$ and is Lipschitz continuous with Lipschitz constant $c_5 > 0$. Further, $w(C_2\sqrt{s(\log p)/n}) > \frac{1}{2}w(0+)$ for large enough n.

Define: $\delta_n^* = \left[\sqrt{s(\log n)/n} + \lambda_n\left(||\boldsymbol{w}_0^*||_2 + C_2 c_5\sqrt{s(\log p)/n}\right)\right]$, where $\boldsymbol{w}^* = (w(|\beta_{0j}|))_{j=1,\ldots,p}$.

Theorem (Ghosh et al., 2020)

Suppose that the assumptions of the previous theorem hold with $\boldsymbol{w} = \boldsymbol{w}^*$ and $\delta_n = \delta_n^*$. Additionally, if Assumptions (A7)-(A8) hold with $\lambda_n s\kappa_n\sqrt{(\log p)/n} \to 0$ then, with probability tending to one, there exists a global minimizer $\widehat{\beta} = (\widehat{\beta}_1^T, \widehat{\beta}_2^T)^T$ of the AW-DPD-LASSO objective function with adaptive weights, such that

$$\left|\left|\widehat{\beta}_1 - \beta_{10}\right|\right|_2 \leq C_1\delta_n^*, \quad \text{and} \quad \widehat{\beta}_2 = \mathbf{0}_{p-s}.$$

- For **SCAD-based weight**, (A8) holds if $\lambda_n > \frac{2}{(a+1)}C_2\sqrt{\frac{s(\log p)}{n}}$
- The weight for **Ad-DPD-LASSO** is not Lipschitz around zero but is so locally on (c, ∞) for any $c > 0$.

Asymptotic Normality of AW-DPD-LASSO

(A6) $\boldsymbol{Z}_n \boldsymbol{\Omega}_n \boldsymbol{Z}_n$ is positive definite, $\lambda_n ||\boldsymbol{w}_0||_2 = O(\sqrt{s/n})$, and $\sqrt{n/s} \min_{1 \le j \le s} |\beta_{j0}| \to \infty$.

(A9) $\min_{1 \le j \le s} |\beta_{0j}| > 2C_2 \sqrt{s(\log p)/n}$ and $w'(|b|) = o(s^{-1}\lambda_n^{-1}(n \log p)^{-1/2})$ for $|b| > \frac{1}{2} \min_{1 \le j \le s} |\beta_{0j}|$.

Theorem (Ghosh et al., 2020)

Suppose that the assumptions of the previous theorems hold along with Assumption (A6) or (A9) according to the weights being fixed or stochastic. Then, with probability tending to one, there exists a global minimizer $\widehat{\beta} = (\widehat{\beta}_1^o, \widehat{\beta}_2)^T$ *of the AW-DPD-LASSO objective function such that*

$$\widehat{\beta}_2 = \mathbf{0}_{p-s}, \quad \text{and} \quad \boldsymbol{u}^T [\boldsymbol{Z}_n^T \boldsymbol{\Omega}_n \boldsymbol{Z}_n]^{-1/2} \boldsymbol{V}_n^{-1} \left[\left(\widehat{\beta}_1^o - \beta_{10} \right) + n\lambda_n \boldsymbol{V}_n^2 \widetilde{\boldsymbol{w}}_0 \right] \xrightarrow{\mathcal{L}} N(0,1),$$

for any arbitrary $\boldsymbol{u} \in \mathbb{R}^s$ *satisfying* $\boldsymbol{u}^T \boldsymbol{u} = 1$, *where* $\widetilde{\boldsymbol{w}}_0 = \left(w_j sign(\beta_{j0}) : j \in S \right)$, *and*

$$\boldsymbol{V}_n = \left[\boldsymbol{X}_1^T E[\boldsymbol{H}_\alpha^{(2)}(\beta_0)] \boldsymbol{X}_1 \right]^{-1/2}, \quad \boldsymbol{Z}_n = \boldsymbol{X}_1 \boldsymbol{V}_n, \quad \boldsymbol{\Omega}_n = Var \left[\boldsymbol{H}_\alpha^{(1)}(\beta_0) \right].$$

For **SCAD-based weight**, (A9) holds if $\min_{1 \le j \le s} |\beta_{0j}| \ge 2a\lambda_n$ (may choose λ_n to let bias $\to 0$).

Corollary: Properties of DPD-LASSO

Corollary

Suppose that Assumption (A1)-(A4) hold true. Then, for a given constant $C_1 > 0$, we have the following results for the DPD-LASSO estimator with regularization parameter $\lambda_n = O(n^{-1/2})$.

a) There exists some $c > 0$ such that, with probability at least $1 - n^{-cs}$, the corresponding oracle estimator $\widehat{\beta}_1^o$ satisfies

$$\left|\left|\widehat{\beta}_1^o - \beta_{10}\right|\right|_2 \leq C_1 \left[\sqrt{s(\log n)/n} + \lambda_n \sqrt{s}\right]. \tag{6}$$

b) If $\sqrt{s(\log n)/n} = o\left(\min_{1 \leq j \leq s} |\beta_{j0}|\right)$, then **sign($\widehat{\beta}_1^o$) = sign($\beta_{10}$)**.

c) If Assumption (A5) holds with $\lambda_n > 2\sqrt{(c+1)\log p/n}$ and $n^{1/2}(\log_2 n)^{5/2} = O(\log p)$, then, with probability at least $1 - O(n^{-cs})$, there exists a global minimizer $\widehat{\beta} = ((\widehat{\beta}_1^o)^T, \widehat{\beta}_2^T)^T$ of the DPD-LASSO objective function such that $\widehat{\beta}_1^o$ satisfies (6) and $\widehat{\beta}_2 = \mathbf{0}_{p-s}$.

d) In addition to the assumptions of item (c), if $\sqrt{n/s}\min_{1 \leq j \leq s} |\beta_{j0}| \to \infty$ and $\boldsymbol{Z}_n \boldsymbol{\Omega}_n Z_n$ is positive definite, then the DPD-LASSO estimator $\widehat{\beta}_1^o$ obtained in item (c) further satisfies the asymptotic normality result, but the associated bias $n\lambda_n \boldsymbol{V}_n \widetilde{\boldsymbol{w}_0}$ is non-diminishing.

Empirical Illustration of AW-DPD-LASSO: Results for Y-outliers

Table: $p = 1000$, $n = 100$, $s = 9$ (sparsely distributed)

Method	MS($\widehat{\beta}$)	TP($\widehat{\beta}$)	TN($\widehat{\beta}$)	MSES($\widehat{\beta}$) (10^{-2})	MSEN($\widehat{\beta}$) (10^{-5})	EE($\widehat{\sigma}$) (10^{-2})	APrB($\widehat{\beta}$) (10^{-2})
LS-LASSO	7.85	0.46	1.00	332.57	56.63	923.77	53.67
Ad-LS-LASSO	5.24	0.42	1.00	282.04	196.13	728.11	46.04
LS-SCAD	26.32	0.63	0.98	253.30	396.77	526.46	44.85
LS-MCP	11.40	0.52	0.99	256.43	316.72	584.97	44.96
LAD-LASSO	25.93	0.85	0.98	144.66	226.31	315.34	36.80
RLARS	20.25	0.79	0.99	90.67	320.57	144.45	26.38
sLTS	51.28	0.93	0.96	83.66	131.04	7.52	22.84
DPD-LASSO $\alpha = 0.1$	8.50	0.74	1.00	65.91	1475.56	348.96	38.15
DPD-LASSO $\alpha = 0.3$	6.94	0.62	1.00	80.41	1681.17	381.90	39.62
DPD-LASSO $\alpha = 0.5$	12.22	0.88	1.00	27.28	618.61	130.50	20.97
DPD-LASSO $\alpha = 0.7$	23.58	0.61	0.98	92.40	1643.77	30.59	36.54
DPD-LASSO $\alpha = 1$	14.03	0.53	0.99	105.81	2131.45	215.98	47.96
DPD-ncv $\alpha = 0.1$	9.10	0.91	1.00	23.00	425.28	127.54	16.34
DPD-ncv $\alpha = 0.3$	9.34	0.98	1.00	4.78	118.05	19.60	6.88
DPD-ncv $\alpha = 0.5$	9.12	0.98	1.00	4.65	110.96	13.06	6.67
DPD-ncv $\alpha = 0.7$	9.21	0.98	1.00	4.47	101.86	10.88	6.64
DPD-ncv $\alpha = 1$	9.22	0.99	1.00	4.49	100.00	10.20	6.62
Ad-DPD-LASSO $\alpha = 0.1$	18.71	0.81	0.99	81.49	835.39	130.08	22.43
Ad-DPD-LASSO $\alpha = 0.3$	10.92	0.99	1.00	6.21	27.67	10.02	6.14
Ad-DPD-LASSO $\alpha = 0.5$	9.98	0.99	1.00	5.81	24.30	5.94	5.50
Ad-DPD-LASSO $\alpha = 0.7$	9.48	0.99	1.00	3.87	11.31	7.30	5.11
Ad-DPD-LASSO $\alpha = 1$	9.96	0.97	1.00	15.38	48.27	8.59	6.66
AW-DPD-LASSO $\alpha = 0.1$	12.98	0.80	0.99	92.47	567.77	169.98	23.06
AW-DPD-LASSO $\alpha = 0.3$	10.84	0.99	1.00	6.36	31.77	9.71	5.91
AW-DPD-LASSO $\alpha = 0.5$	12.72	0.99	1.00	3.84	15.54	7.87	4.91
AW-DPD-LASSO $\alpha = 0.7$	10.80	0.99	1.00	3.69	13.75	8.37	5.14
AW-DPD-LASSO $\alpha = 1$	11.51	0.96	1.00	17.81	57.74	8.38	8.38

Empirical Illustration of AW-DPD-LASSO: Results for X-outliers

Table: $p = 1000$, $n = 100$, $s = 9$ (sparsely distributed)

Method	MS($\widehat{\boldsymbol{\beta}}$)	TP($\widehat{\boldsymbol{\beta}}$)	TN($\widehat{\boldsymbol{\beta}}$)	MSES($\widehat{\boldsymbol{\beta}}$) ($10^{-2}$)	MSEN($\widehat{\boldsymbol{\beta}}$) ($10^{-5}$)	EE($\widehat{\sigma}$) (10^{-2})	APrB($\widehat{\boldsymbol{\beta}}$) ($10^{-2}$)
LS-LASSO	7.85	0.46	1.00	332.57	56.63	923.77	53.67
Ad-LS-LASSO	9.00	1.00	1.00	1.52	0.00	31.56	4.85
LS-SCAD	26.32	0.63	0.98	253.30	396.77	526.46	44.85
LS-MCP	11.40	0.52	0.99	256.43	316.72	584.97	44.96
LAD-LASSO	25.93	0.85	0.98	144.66	226.31	315.34	36.80
RLARS	20.25	0.79	0.99	90.67	320.57	144.45	26.38
sLTS	51.28	0.93	0.96	83.66	131.04	7.52	22.84
DPD-LASSO $\alpha = 0.1$	10.33	1.00	1.00	4.58	114.30	58.74	10.80
DPD-LASSO $\alpha = 0.3$	8.71	0.90	1.00	27.94	700.75	211.92	26.32
DPD-LASSO $\alpha = 0.5$	13.08	0.99	1.00	3.14	96.10	35.47	9.83
DPD-LASSO $\alpha = 0.7$	21.04	0.79	0.99	56.38	945.35	31.38	24.17
DPD-LASSO $\alpha = 1$	14.30	0.69	0.99	73.45	1515.09	164.59	37.40
DPD-ncv $\alpha = 0.1$	9.00	1.00	1.00	0.14	2.64	5.58	3.88
DPD-ncv $\alpha = 0.3$	9.00	1.00	1.00	0.17	3.09	10.29	3.89
DPD-ncv $\alpha = 0.5$	9.00	1.00	1.00	0.19	3.38	13.85	3.94
DPD-ncv $\alpha = 0.7$	9.00	1.00	1.00	0.23	3.81	16.61	3.99
DPD-ncv $\alpha = 1$	9.01	1.00	1.00	0.28	4.14	19.80	4.10
Ad-DPD-LASSO $\alpha = 0.1$	11.20	1.00	1.00	0.50	2.62	5.91	4.20
Ad-DPD-LASSO $\alpha = 0.3$	9.10	1.00	1.00	0.64	0.04	3.04	4.15
Ad-DPD-LASSO $\alpha = 0.5$	9.58	1.00	1.00	0.66	0.45	4.65	4.13
Ad-DPD-LASSO $\alpha = 0.7$	9.16	1.00	1.00	0.79	0.11	5.30	4.17
Ad-DPD-LASSO $\alpha = 1$	9.18	1.00	1.00	1.06	0.39	8.68	4.02
AW-DPD-LASSO $\alpha = 0.1$	10.68	1.00	1.00	0.34	0.28	3.92	3.58
AW-DPD-LASSO $\alpha = 0.3$	9.42	1.00	1.00	0.36	0.00	3.88	3.76
AW-DPD-LASSO $\alpha = 0.5$	10.70	1.00	1.00	0.43	0.30	5.46	3.80
AW-DPD-LASSO $\alpha = 0.7$	9.80	1.00	1.00	0.50	0.20	6.67	3.94
AW-DPD-LASSO $\alpha = 1$	9.58	1.00	1.00	0.67	0.09	9.56	4.02

Empirical Illustration of AW-DPD-LASSO: Results for No Outliers

Table: $p = 1000$, $n = 100$, $s = 9$ (sparsely distributed)

Method	MS($\widehat{\boldsymbol{\beta}}$)	TP($\widehat{\boldsymbol{\beta}}$)	TN($\widehat{\boldsymbol{\beta}}$)	MSES($\widehat{\boldsymbol{\beta}}$) ($10^{-2}$)	MSEN($\widehat{\boldsymbol{\beta}}$) ($10^{-5}$)	EE($\widehat{\sigma}$) (10^{-2})	APrB($\widehat{\boldsymbol{\beta}}$) ($10^{-2}$)
LS-LASSO	21.27	1.00	0.99	2.52	1.52	34.91	6.31
Ad-LS-LASSO	9.00	1.00	1.00	1.49	0.00	31.30	5.16
LS-SCAD	9.03	1.00	1.00	0.36	0.00	19.62	4.23
LS-MCP	9.04	1.00	1.00	0.36	0.00	19.60	4.22
LAD-LASSO	16.70	1.00	0.99	7.55	2.63	53.97	9.17
RLARS	12.70	1.00	1.00	0.48	2.82	7.65	4.49
sLTS	61.80	0.79	0.94	219.62	300.30	6.79	46.08
DPD-LASSO $\alpha = 0.1$	10.37	1.00	1.00	4.59	114.54	58.75	11.10
DPD-LASSO $\alpha = 0.3$	8.73	0.90	1.00	27.95	701.22	211.92	25.86
DPD-LASSO $\alpha = 0.5$	13.13	0.99	1.00	3.13	95.99	35.29	10.61
DPD-LASSO $\alpha = 0.7$	22.12	0.73	0.98	70.05	1201.08	25.81	29.53
DPD-LASSO $\alpha = 1$	14.40	0.69	0.99	73.96	1500.41	163.41	37.38
DPD-ncv $\alpha = 0.1$	9.00	1.00	1.00	0.14	2.64	5.58	4.20
DPD-ncv $\alpha = 0.3$	9.00	1.00	1.00	0.17	3.08	10.28	4.21
DPD-ncv $\alpha = 0.5$	9.00	1.00	1.00	0.19	3.36	13.86	4.23
DPD-ncv $\alpha = 0.7$	9.00	1.00	1.00	0.23	3.82	16.63	4.25
DPD-ncv $\alpha = 1$	9.01	1.00	1.00	0.26	4.12	19.82	4.31
Ad-DPD-LASSO $\alpha = 0.1$	11.34	1.00	1.00	0.49	2.79	6.18	4.24
Ad-DPD-LASSO $\alpha = 0.3$	9.10	1.00	1.00	0.64	0.04	3.04	4.33
Ad-DPD-LASSO $\alpha = 0.5$	9.58	1.00	1.00	0.66	0.45	4.65	4.39
Ad-DPD-LASSO $\alpha = 0.7$	9.16	1.00	1.00	0.79	0.11	5.29	4.31
Ad-DPD-LASSO $\alpha = 1$	9.18	1.00	1.00	1.06	0.38	8.68	4.36
AW-DPD-LASSO $\alpha = 0.1$	10.68	1.00	1.00	0.34	0.28	3.92	3.78
AW-DPD-LASSO $\alpha = 0.3$	9.62	1.00	1.00	0.36	0.00	3.93	3.92
AW-DPD-LASSO $\alpha = 0.5$	10.70	1.00	1.00	0.43	0.30	5.46	4.00
AW-DPD-LASSO $\alpha = 0.7$	9.82	1.00	1.00	0.50	0.20	6.68	4.04
AW-DPD-LASSO $\alpha = 1$	9.60	1.00	1.00	0.66	0.09	9.56	3.99

Outline

1. Robust Adaptive Procedures
2. **Robust Variable Screening**
3. Stability of the Set of Selected Variables
4. Conclusion

Variable Screening for Extremely High-dimensional Data

Standard **linear regression model** (LRM): $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$,

Simultaneous estimation and variable selection often fail (computationally) if $p \ggg n$

Sure Independence Screnning (SIS) of variables (Fan and Lv, 2008)

- Reduce model size (p) by an initial variable screening; then apply regularized procedure!
- **Rank the covariates by their absolute correlation values with response** and select top d!
- **Sure Screening property**: Selected covariate list contains all truly important variable, asymptotically with probability tending to one.

Alternative Formulation: Marginal Regression Approach

- Given j-th covariate X_j, we consider the j-th marginal model $y_i = \gamma_j + \beta_j x_{ij} + \epsilon_{ij}$, where the ϵ_{ij}s are IID for $i = 1, \ldots, n$, each having distribution $N(0, \sigma_j^2)$.
- Estimate $\boldsymbol{\theta}_j = (\gamma_j, \beta_j, \sigma_j)^T$ by usual MLE or OLS based methods, say, $(\widehat{\gamma}_j, \widehat{\beta}_j, \widehat{\sigma})$.
- If all covariates are standardized, ranking them in order of (absolute) correlation with the response is equivalent to ordering the absolute value of estimated marginal slopes $|\widehat{\beta}_j|$.

SIS for GLM (Fan and Song, 2010)

The marginal approach can directly be extended for GLM ranking covaraites it terms of the absolute value of the MLE of their marginal slopes (Sure Screening property holds).

Robust Screening Procedures for Linear Regression Model

Standard **linear regression model** (LRM): $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$,

Usual SIS is extremely Non-robust under data contamination!

Approches using Robust Correlation Measures

- **GK-SIS** (Gather and Guddat, 2008): A robust correlation measure proposed by Gnanadesikan and Kettenring (1972).
- **Rank-SIS** (Li et al., 2012a): non-parametric rank correlation.
- **dCor-SIS** (Li et al., 2012b; Wang et al., 2017): a distance based correlation measure from Szekely et al. (2007).
- **MCP-SIS** (Mu and Xiong, 2014): A robust measure of association, namely the median of component-wise products (MCPs).
- **BW-SIS** (Mu and Xiong, 2014): The bivariate winsorized (BW) correlation estimator of Khan et al. (2007).

A Robust and Efficient Variable Screening Procedure

Standard **linear regression model** (LRM): $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$,

Marginal Regression Model

Given j-th covariate X_j, we consider the j-th marginal model $y_i = \gamma_j + \beta_j x_{ij} + \epsilon_{ij}$, where the ϵ_{ij}s are IID for $i = 1, \ldots, n$, each having distribution $N(0, \sigma_j^2)$.

Minimum DPD Estimation for Marginal Regression (Ghosh and Basu, 2013)

$$\widehat{\boldsymbol{\theta}}_j^M = (\widehat{\gamma}_j^{M\alpha}, \widehat{\beta}_j^{M\alpha}, \widehat{\sigma}_j^{M\alpha}) = \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} l_\alpha \left(y_i, \gamma_j + \beta_j x_{ij}, \sigma\right).$$

where

$$l_\alpha(y, \eta, \sigma) = \frac{1}{\sigma^\alpha (2\pi)^{\alpha/2}} \left(\frac{1}{\sqrt{1+\alpha}} - \frac{1+\alpha}{\alpha} e^{-\frac{\alpha(y-\eta)^2}{\sigma^2}} \right) + \frac{1}{\alpha}.$$

Variable Screening using MDPDE (Ghosh and Thoresen, 2021)

- Rank covaraites in terms of absolute value of the MDPDE of their marginal slopes, $|\widehat{\beta}_j^{M\alpha}|$.
- As $\alpha \downarrow 0$, the MDPDE coincides with the MLE, and we get the usual (non-robust) SIS.
- At $\alpha > 0$, we get a robust generalization of SIS that remains stable under data contamination.

The DPD-based Screening Procedure

Algorithm: DPD-SIS(α)

1. **Input:** n-vector of responses $\boldsymbol{y}$; $n \times p$ matrix of (standardized) covariates $\boldsymbol{X}$; model size d.

2. For each $j = 1, \ldots, p$, compute the marginal MDPDE $\widehat{\beta}_j^{M\alpha}$ as

$$\widehat{\boldsymbol{\theta}}_j^M = (\widehat{\gamma}_j^{M\alpha}, \widehat{\beta}_j^{M\alpha}, \widehat{\sigma}_j^{M\alpha}) = \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} l_\alpha \left(y_i, \gamma_j + \beta_j x_{ij}, \sigma \right).$$

3. Sort $|\widehat{\beta}_j^{M\alpha}|$ in decreasing order for $j = 1, \ldots, p$. Set $r_k = j$, if $|\widehat{\beta}_j^{M\alpha}|$ has rank k, for each k.

4. Construct the estimated model set $\widehat{\mathcal{M}}_\alpha(d) = \{r_1, \ldots, r_d\}$, with indices corresponding to the top d values of (absolute) marginal MDPDEs.

5. Run a robust penalized regression model (low or moderate dimensional) with the covariates selected in $\widehat{\mathcal{M}}_\alpha(d)$ to obtain an estimated coefficient vector, say $\widehat{\boldsymbol{\beta}}_d = (\widehat{\beta}_{d0}, \widehat{\beta}_{dr_1}, \ldots, \widehat{\beta}_{dr_d})^T$.

   (We suggest to use the DPD based method with the same α, which also gives an estimate $\widehat{\sigma}^2$ of the overall model error variance σ^2.)

6. **Output:** The final estimated model $\widehat{\mathcal{M}} = \left\{ 1 \le k \le d : \widehat{\beta}_{dr_k} \neq 0 \right\}$ along with the parameter estimates $\widehat{\boldsymbol{\beta}}_d$ (and the estimate $\widehat{\sigma}^2$ of σ^2, if available).

Iterative DPD-SIS: The DPD-ISIS(α)

1. **Input:** n-vector of responses $\boldsymbol{y}$; $n \times p$ matrix of (standardized) covariates $\boldsymbol{X}$; model size d.
2. Set $i = 1$, $\boldsymbol{y}^{(1)} = \boldsymbol{y}$ and index set of available covariates as $\mathcal{W}_1 = \{1, \ldots, p\}$
3. **DPD-SIS with model size d':**
   1. For each $j \in \mathcal{W}_i$, compute the marginal MDPDE $\widehat{\beta}_j^{M\alpha}$ and order them (according to their absolute value) to construct the estimated model set $\widehat{\mathcal{M}}_\alpha^{(i)} = \{r_1, \ldots, r_{d'}\}$.
4. Run any suitable (fast) robust penalized regression model with the main response $\boldsymbol{y}$ and the covariates selected in $\cup_{k=1}^{i} \widehat{\mathcal{M}}_\alpha^{(k)}$ to get estimated coefficient vector $\widehat{\boldsymbol{\beta}}^{(i)} = (\widehat{\beta}_0^{(i)}, \widehat{\beta}_{j_1}^{(i)}, \ldots, \widehat{\beta}_{j_{k_i}}^{(i)})^T$.

   Denote $\mathcal{A}_i = \left\{ j_a : \widehat{\beta}_{j_a}^{(i)} \neq 0, a = 1, \ldots, k_i \right\} \subset \mathcal{W}_i$.
5. If a specified stopping criterion is satisfied, go to step 8. Otherwise go to Step 6.
6. Compute the residuals $\boldsymbol{r}^{(i)} = \boldsymbol{y} - \boldsymbol{X}_{\mathcal{A}_i} \widehat{\boldsymbol{\beta}}^{(i)}$.
7. Set $\boldsymbol{y}^{(i+1)} = \boldsymbol{r}^{(i)}$ and the index set of available covariates as $\mathcal{W}_{i+1} = \mathcal{W}_i \setminus \mathcal{A}_i$.
   Change i to $i + 1$ and go to Step 3.
8. Final Stage Model: Run a robust penalized regression model (low or moderate dimensional) with the covariates selected in $\mathcal{A}_i$ to get estimated coefficient vector, say $\widehat{\boldsymbol{\beta}}_d = (\widehat{\beta}_{d0}, \widehat{\beta}_{dr_1}, \ldots, \widehat{\beta}_{dr_d})^T$.
9. **Output:** The final estimated model $\widehat{\mathcal{M}} = \left\{ 1 \leq k \leq d : \widehat{\beta}_{dr_k} \neq 0 \right\}$ along with the parameter estimates $\widehat{\boldsymbol{\beta}}_d$ (and the estimate $\widehat{\sigma}^2$ of σ^2, if available).
10. **Stopping Criteria**: $|\mathcal{A}_i| \geq d$ or $\mathcal{A}_i = \mathcal{A}_{i-1}$ or a fixed number of iteration!

The DPD-SIS for GLM

Generalized linear model (GLM): Given a covariate value $\mathbf{X} = \boldsymbol{x}$, the response variable Y has density

$$f(y; \boldsymbol{x}^T\beta) = \exp\{y\theta - b(\theta) + c(y)\}, \quad \text{with} \quad E[Y|\boldsymbol{x}] = b'(\theta) = g^{-1}(\boldsymbol{x}^T\beta), \tag{7}$$

where $b(\cdot)$ and $c(\cdot)$ are known functions, g is a known monotone differentiable link function, and the canonical parameter θ is defined via the linear predictor $\eta = \boldsymbol{x}^t\beta$.

DPD Loss Function (Ghosh and Basu, 2016)

$$\widehat{\beta}_j^{M\alpha} = \left(\widehat{\beta}_{j0}^{M\alpha}, \widehat{\beta}_j^{M\alpha}\right) = \arg\min_{\beta_{j0},\beta_j} \frac{1}{n}\sum_{i=1}^{n} l_\alpha\left(y_i, \beta_{j0} + \beta_j x_{ij}\right),$$

where $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^T$ for each $i = 1, \ldots, n$, and

$$l_\alpha(y, \theta) = \int f(s;\theta)^{1+\alpha} ds - \left(1 + \frac{1}{\alpha}\right) f(y;\theta)^\alpha + \frac{1}{\alpha}.$$

DPD-SIS for GLM (Ghosh et al., 2021)

- Rank covaraites in terms of absolute value of the MDPDE of their marginal slopes, $|\widehat{\beta}_j^{M\alpha}|$.

Theoretical Justification: Population-level Results

- The population version (functional) of the marginal MDPDE $\widehat{\beta}_j^{M\alpha}$ is given by

$$\boldsymbol{\beta}_j^{M\alpha} = (\beta_{j0}^{M\alpha}, \beta_j^{M\alpha}) = \arg\min_{\beta_{j0},\beta_j} E\left[l_\alpha\left(Y, \beta_{j0} + \beta_j X_j\right)\right].$$

- Assume: Covariates are centered and standardized.
- Define $B_\alpha(v(\boldsymbol{x})) = b'(\boldsymbol{x}^T\boldsymbol{\beta}_0) - E[\psi_\alpha(Y, v(\boldsymbol{x}))|\boldsymbol{X} = \boldsymbol{x}]$, where $\psi_\alpha(y,\theta) = \frac{\partial}{\partial\theta} l_\alpha(y,\theta)$.

Theorem (Ghosh et al., 2021)

For a given $\alpha \geq 0$, and for any $j = 1, \ldots, p$, the marginal MDPDE functional $\beta_j^{M\alpha} = 0$ if and only if

$$Cov(b'(\boldsymbol{X}^T\boldsymbol{\beta}_0), X_j) = Cov(Y, X_j) = 0.$$

Theorem (Ghosh et al., 2021)

Given any $\alpha \geq 0$, along with the assumptions of the above theorem on $B_\alpha(\cdot)$, let us additionally assume that either $B'_\alpha(\cdot)$ is bounded or $G_\alpha(|x|) = \sup_{|u|\leq|x|} |B_\alpha(u)|$ satisfies

$$E[G(a|X_j|)|X_j|I(|X_j| \geq n^\eta)] \leq dn^{-\kappa},$$

for all $j \in \mathcal{M}_0$ and constants $a, d > 0$, $\eta \in (0, \kappa)$. If there exists a constant $c_1 > 0$ such that $\left|Cov(b'(\boldsymbol{X}^T\boldsymbol{\beta}_0), X_j)\right| \geq c_1 n^{-\kappa}$ for all $j \in \mathcal{M}_0$, then we have $\min_{j\in\mathcal{M}_0} |\beta_j^{M\alpha}| \geq c_2 n^{-\kappa}$ for $c_2 > 0$.

Sure Screening Property of the DPD-SIS

Theorem (Exponential Consistency of Marginal MDPDEs)

Suppose that (A1)–(A5) hold for a given $\alpha \geq 0$. Then, for any $t > 0$,

$$P\left(\sqrt{n}\left|\widehat{\beta}_j^{M\alpha} - \beta_j^{M\alpha}\right| \geq \frac{16k_n^{(\alpha)}}{V}(1+t)\right) \leq e^{-\frac{2t^2}{K_n^2}} + nm_1 e^{-m_0 K_n^{\tau}}, \quad j = 1, \ldots, p, \tag{8}$$

where $k_n = k_n^{(\alpha)} = (1+\alpha)\left[\frac{m_0}{m_3}K_n^2 L_\alpha + |b'(K_n B + B)|L_\alpha + \xi_\alpha(K_n B + B)\right]$.

Theorem (Ghosh et al., 2021)

Let Assumptions (A1)–(A5) hold for a given $\alpha \geq 0$ and $\frac{n^{1-2\kappa}}{(k_n K_n)^2} \to \infty$ as $n \to \infty$.

(a) *For any given $c_3 > 0$, there exists $C > 0$ such that*

$$P\left(\max_{1\leq j\leq p}|\widehat{\beta}_j^{M\alpha} - \beta_j^{M\alpha}| \geq c_3 n^{-\kappa}\right) \leq pR_n, \qquad R_n = \left[e^{-\frac{n^{1-2\kappa}C}{(k_n K_n)^2}} + nm_1 e^{-m_0 K_n^{\tau}}\right].$$

(b) *Under the additional assumptions ensuring the population-level results, then taking $\gamma_n = c_4 n^{-\kappa}$ with $c_4 \leq c_2/2$, we have*

$$P\left(\widehat{\mathcal{M}}(\gamma_n) \supset \mathcal{M}_0\right) \geq 1 - sR_n.$$

(c) *If additionally Assumptions (A6)–(A7) hold, taking $\gamma_n = c_4 n^{-2\kappa}$, $c_4 > 0$, we get*

$$P\left(|\widehat{\mathcal{M}}(\gamma_n)| \leq O(n^{2\kappa}\Lambda_{\max}(\Sigma))\right) \geq 1 - pR_n.$$

* Required Assumptions

(A1) The GLM is such that f^{α} is bounded (say, by $L_{\alpha} > 0$) and the function $b''(\cdot)$ is continuous and positive. Also, $|\xi_{\alpha}(\cdot)|$ is non-decreasing.

(A2) For all $\boldsymbol{\beta}_j \in \mathcal{B}$, there exists some constant $V > 0$ such that $\Lambda_{\min}\left[\boldsymbol{J}_{j,\alpha}(\boldsymbol{\beta}_j)\right] \geq V$ uniformly over $j = 1, \ldots, p$.

(A3) $\boldsymbol{K}_{j,\alpha}(\boldsymbol{\beta}_j^{M\alpha})$ is finite and positive definite for each $j = 1, \ldots, p$. Also, the norm $||\boldsymbol{K}_{j,\alpha}(\boldsymbol{\beta}_j)||_{\mathcal{B}} = \sup_{\boldsymbol{\beta}_j \in \mathcal{B}, ||\boldsymbol{u}||=1} ||\boldsymbol{K}_{j,\alpha}(\boldsymbol{\beta}_j)^{1/2}\boldsymbol{u}||$ is bounded from above for each j.

(A4) There exist an $\epsilon_1 > 0$ and a large constant $K_n > 0$, such that

$$\sup_{\boldsymbol{\beta}_j \in \mathcal{B}: ||\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^{M\alpha}|| \leq \epsilon_1} E\left[|B_{\alpha}(\boldsymbol{X}_j^T\boldsymbol{\beta}_j)|||\boldsymbol{X}_j||_2 I(|X_j| > K_n)\right] \leq o\left(\frac{1}{n}\right), \qquad \text{for all } j = 1, 2, \ldots, p.$$

(A5) The distribution of the covariate X_j is such that, for sufficiently large $t > 0$ and some positive constants m_0, m_1, m_2, m_3 and τ, we have $P(|X_j| > t) = (m_1 - m_2)e^{-m_0 t^{\tau}}$, for $j = 1, 2, \ldots, p$, and

$$E\left[\exp\left(b(\boldsymbol{X}^T\boldsymbol{\beta}_0 + m_3) - b(\boldsymbol{X}^T\boldsymbol{\beta}_0)\right)\right] + E\left[\exp\left(b(\boldsymbol{X}^T\boldsymbol{\beta}_0 - m_3) - b(\boldsymbol{X}^T\boldsymbol{\beta}_0)\right)\right] \leq m_2.$$

(A6) $Var(\boldsymbol{X}^T\boldsymbol{\beta}_0)$ is bounded both from below and above by finite positive constants.

(A7) Either $b''(\cdot)$ is bounded or $\widetilde{\boldsymbol{X}} = (X_1, \ldots, X_p)^T$ follows an elliptically contoured distribution with variance Σ_1 and $\left|E\left[b'(\boldsymbol{X}^T\boldsymbol{\beta}_0)(\boldsymbol{X}^T\boldsymbol{\beta}_0 - \beta_{00})\right]\right|$ is bounded.

Robust Conditional Variable Screening

Conditional DPD-SIS

- It may be known before hand that a few variables are important, say $\boldsymbol{X}_{\mathcal{C}} = (X_1, \ldots, X_q)$.
- In presence of variables in $\boldsymbol{X}_{\mathcal{C}}$, we need to screen from remaining $\boldsymbol{X}_{\mathcal{D}} = (X_{q+1}, \ldots, X_p)$
- For a given $\alpha \geq 0$, we may choose the variables from $\boldsymbol{X}_{\mathcal{D}}$ based on the marginal MDPDEs:

$$\widehat{\boldsymbol{\beta}}_{\mathcal{C}j}^{M\alpha} = \left(\widehat{\boldsymbol{\beta}}_{\mathcal{C}j1}^{M\alpha}, \widehat{\beta}_j^{M\alpha}\right) = \arg\min_{\boldsymbol{\beta}_{\mathcal{C}}, \beta_j} \frac{1}{n}\sum_{i=1}^{n} l_\alpha\left(y_i, \boldsymbol{x}_{i\mathcal{C}}^T\boldsymbol{\beta}_{\mathcal{C}} + \beta_j x_{ij}\right), \quad j = q+1, \ldots, p.$$

- Given a pre-defined threshold γ_n, we select $\widehat{M}_\alpha(\gamma_n|\mathcal{D}) = \left\{q+1 \leq j \leq p : \left|\widehat{\beta}_j^{M\alpha}\right| \geq \gamma_n\right\}$.

Theorem (Ghosh et al., 2021)

For a given $\alpha \geq 0$ and any $j \in \mathcal{D}$, the (conditional) marginal MDPDE functional $\beta_j^{M\alpha} = 0$ if and only if $Cov_L(Y, X_j|\boldsymbol{X}_{\mathcal{C}}) := E\left[\left(Y - E_L[Y|\boldsymbol{X}_{\mathcal{C}}]\right)\left(X_j - E_L[X_j|\boldsymbol{X}_{\mathcal{C}}]\right)\right] = 0$.

Theorem (Ghosh et al., 2021)

Given any $\alpha \geq 0$, suppose that $E[m_{\alpha,j}X_j^2] \leq c_2$ uniformly in $j \in \mathcal{D}$, for some constant c_2. If there exist constants $c_1 > 0, \kappa < -1/2$ such that $\left|Cov_L(Y, X_j|\boldsymbol{X}_{\mathcal{C}})\right| \geq c_1 n^{-\kappa}$ for all $j \in \mathcal{M}_{0\mathcal{D}}$, then we have $\min_{j \in \mathcal{M}_{0\mathcal{D}}} |\beta_j^{M\alpha}| \geq c_3 n^{-\kappa}$, for another constant $c_3 > 0$.

Sure Screening Property of Conditional DPD-SIS

Theorem (Ghosh et al., 2021)

Suppose that, for a given $\alpha \geq 0$, Assumptions (A1)–(A5) hold with β_j and $\beta_j^{M\alpha}$ replaced by $\beta_{\mathcal{C}j} \in \mathbb{R}^{q+1}$ and $\beta_{\mathcal{C}j}^{M\alpha}$, respectively, in (A2)–(A4). Also, let $\frac{n^{1-2\kappa}}{(k_n K_n)^2} \to \infty$ as $n \to \infty$.

(a) *For any given $c_3 > 0$, there exists $C > 0$ such that*

$$P\left(\max_{q+1 \leq j \leq p} |\widehat{\beta}_j^{M\alpha} - \beta_j^{M\alpha}| \geq c_3 n^{-\kappa}\right) \leq dR_n, \qquad R_n = \left[e^{-\frac{n^{1-2\kappa} C}{(k_n K_n)^2}} + n m_1 e^{-m_0 K_n^{\tau}}\right].$$

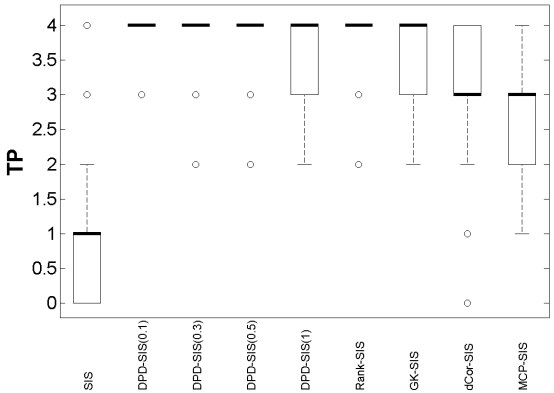
(b) *If additionally the assumptions of previous theorem hold, then taking $\gamma_n = c_4 n^{-\kappa}$ with $c_4 \leq c_2/2$, we have*

$$P\left(\widehat{\mathcal{M}}(\gamma_n) \supset \mathcal{M}_0\right) \geq 1 - sR_n.$$

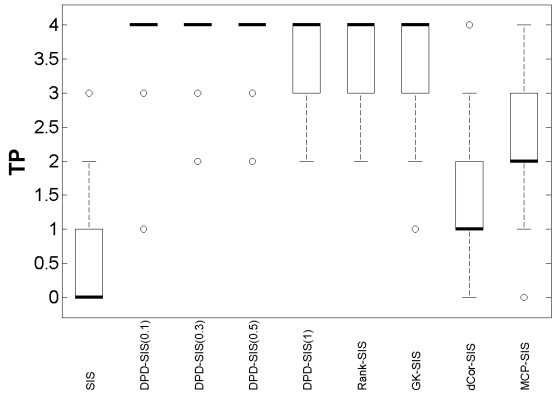
(c) *If additionally Assumptions (A6)–(A8) hold, taking $\gamma_n = c_4 n^{-2\kappa}$, $c_4 > 0$, we get*

$$P\left(|\widehat{\mathcal{M}}(\gamma_n)| \leq O\left(n^{2\kappa} \Lambda_{\max}\left(\Sigma_{\mathcal{D}|\mathcal{C}} + \boldsymbol{Z}\boldsymbol{Z}^T\right)\right)\right) \geq 1 - dR_n.$$

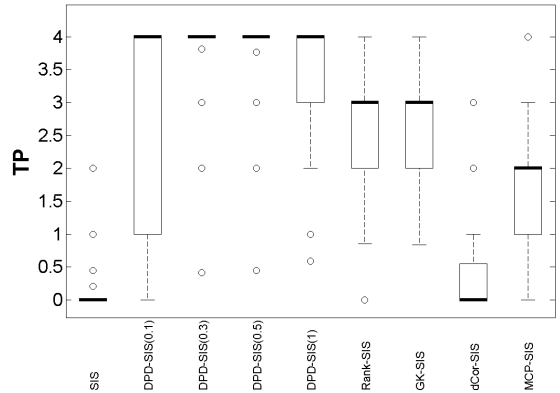
Empirical Illustration: DPD-SIS in Linear Regression



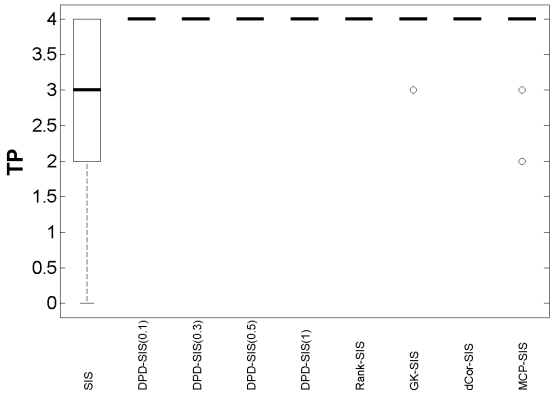
(a) Set 1; 5% contamination



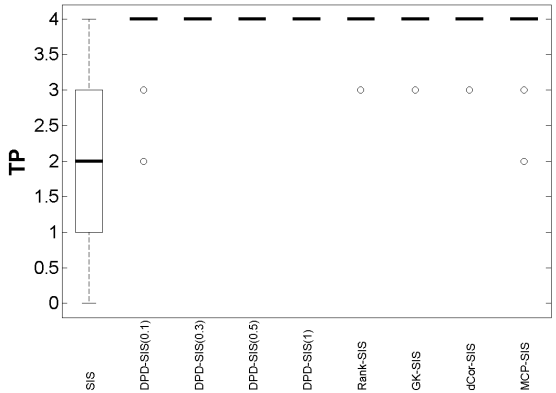
(b) Set 1; 10% contamination



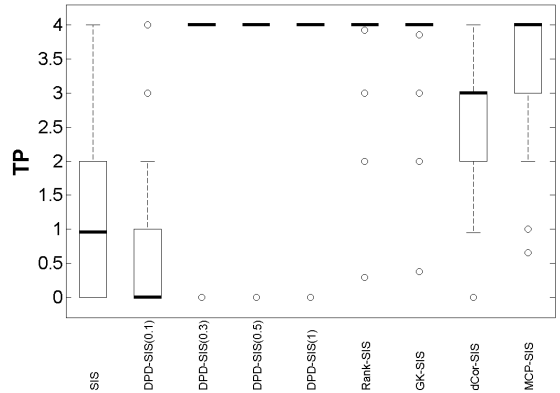
(c) Set 1; 20% contamination



(d) Set 2; 5% contamination



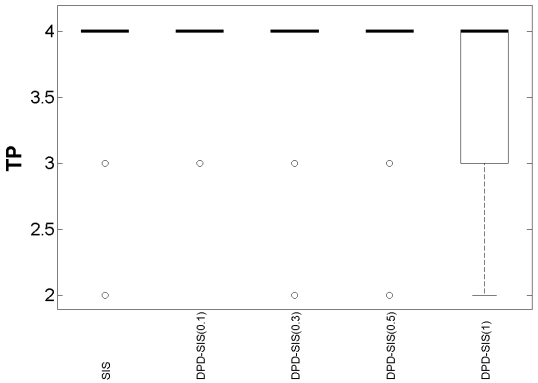
(e) Set 2; 10% contamination



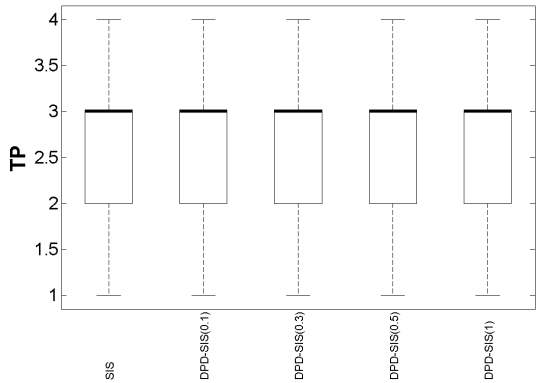
(f) Set 2; 20% contamination

Figure: Box-Plots of the true-positives (TP) obtained by different SIS approaches with target model size $d = n - 1$ for independent covariates (Set 1) and AR correlated covariates (Set 2) with $n = 100$ and moderate signal strength under contamination in data.

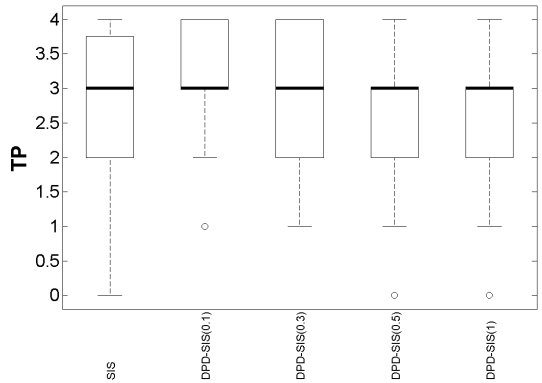
Empirical Illustration: DPD-SIS in Logistic Regression



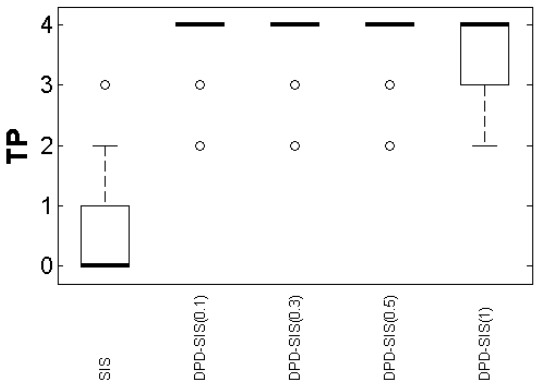
(a)



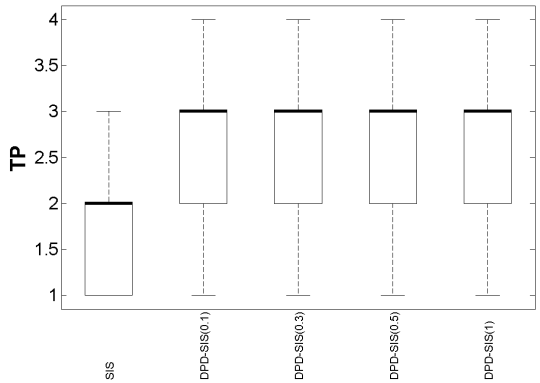
(b)



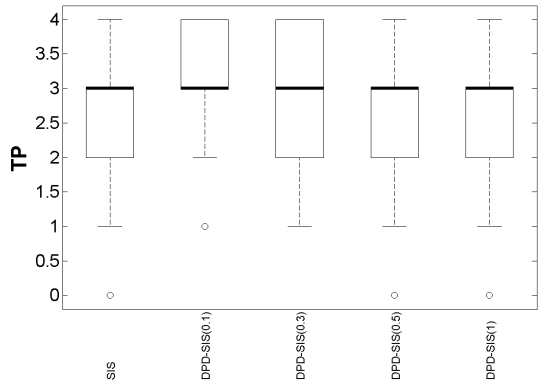
(c)



(d) Independent Covariates ($\rho = 0$), Strong Signal (non-zero values of β_0 are all 5)



(e) Independent Covariates, Weaker Signal $\beta_{0S} = (1, 2, 3, 1, 5)$



(f) Dependent Covariates ($\rho = 0.3$), Strong Signal (non-zero values of β_0 are all 5)

Figure: The box-plots of true-positives selected by the DPD-SIS at different α for different simulation set-ups with pure data (left panel) and 10% contaminated data (right panel). Here non-zero coefficients of β_0 are in positions 1 (intercept), 2, 6, 26 and 126.

Outline

1. Robust Adaptive Procedures
2. Robust Variable Screening
3. **Stability of the Set of Selected Variables**
4. Conclusion

Stability Selection

In real-life application with high-dimensional data, a set of variable often gets highly affected by sampling fluctuation and leads to false positives (and negatives)!

Stability Selection (Meinshausen and Buhlmann, 2010)

- To remove sampling fluctuation from the results and control false discovery!
- Basic idea: Repeat the same penalized estimation method on several random subsamples of size $[n/2]$ from the given data
- Measure the selection stability of each variable by looking at the proportion of its selection among all replications!
- Choose the variables with higher proportion of selection stability, say greater than γ.

This procedure guarantees **an upper bound on the expected number (V) of falsely selected variables**!

$$E(V) \leq \frac{1}{2\gamma - 1} \frac{E[|\widehat{S}|]}{p}, \quad \text{for } \gamma \in (0.5, 1).$$

Stability Selection: An Improved version

Complementary pairs stability selection (Shah and Samworth, 2013)

- Take a random subsample $\mathcal{A}_k$ of size $[n/2]$ from the given data and consider another *complementary sample* as $\widetilde{\mathcal{A}}_k = \{1, 2, \ldots, n\} \setminus \mathcal{A}_k$.
- Apply the chosen penalized estimation procedure on both the subsamples $\mathcal{A}_k$ and $\widetilde{\mathcal{A}}_k$ and let the selected set of variables are $\widehat{S}_{k1}$ and $\widehat{S}_{k2}$, respectively.
- Repeat the process for $k = 1, \ldots, B$ (large number) by randomly choosing a different $\mathcal{A}_k$ each time.
- For j-th variable, compute its **Stability Percentage**

$$SP = \frac{1}{2B} \sum_{k=1}^{B} \left\{ I(j \in \widehat{S}_{k1}) + I(j \in \widehat{S}_{k2}) \right\} \times 100\%.$$

- Finally select the variables having **SP** $\geq \gamma$, a pre-specified threshold.

- Improved bound on the the expected number of false discovery under weaker assumptions!
- Additional bound on the number of important variables excluded.
- These bounds can be significantly sharpened under mild shape restrictions (e.g. unimodality or r-concavity) on the distribution of the stability proportion.

Outline

1. Robust Adaptive Procedures
2. Robust Variable Screening
3. Stability of the Set of Selected Variables
4. **Conclusion**

Summary

- We discussed robust adaptive procedures with more details for the Adaptive DPD-LASSO
- We discussed SIS for variable screening with extremely high-dimensional real-life data and discussed its robust extensions using robust MDPDE based approach.
- The DPD-SIS has been explored in GLM and its sure screening property is discussed, along with the conditional version of DPD-SIS
- We have discussed why it is important to perform Stability Selection in any real-life high-dimensional data analysis
- Numerical illustrations are provided for comparison of different adaptove procedures with DPD and different robust variable screening procedures.

Useful R-packages for Robust High-dimensional Data Analysis

Package	Description
`robustHD`	RLARS and sLTS for LRM.
`enetLTS`	sLTS and enetLTS for LRM and logistic regression.
`pense`	Penalized S-estimator for LRM with l_1 and elastic-net penalties.
`flare`	LASSO, LAD-LASSO, Dantzig Selector and LASSO with l_q loss for LRM.
`gamreg`	Minimum penalized LDPD estimator with l_1 penalty.
`awDPDlasso`[1]	AW-DPD-LASSO for linear and logistic regression models.
`SIS`	Sure Independence Screening and iterative Sure Independence Screening.
`dpdSIS`[2]	DPD based robust Sure Independence Screening.
`stabs`	Stability selection with any robust or non-robust procedure.

[1] https://github.com/MariaJaenada/awDPDlasso

[2] https://github.com/abhianik/dpdSIS

References

- Chang, L., Roberts, S., and Welsh, A. (2018). Robust lasso regression using Tukey's biweight criterion. *Technometrics*, **60(1)**, 36–47.
- Fan, J., Fan, Y., and Barut, E. (2014). Adaptive robust variable selection. *Ann. Stat.*, **42(1)**, 324.
- Fan, J., and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. Royal Stat. Soc. B* , **70(5)**, 849–911.
- Fan, J., and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann Stat*, **38(6)**, 3567–3604.
- Gather, U. and Guddat, C. (2008) Comment on "Sure Independence Screening for Ultrahigh Dimensional Feature Space" by Fan, JQ and Lv, J. *J Royal Stat Soc B*, **70**, 893–895.
- Ghosh, A., and Basu, A. (2013). Robust estimation for independent non-homogeneous observations using density power divergence with applications to linear regression. *Electron. J. Stat.*, **7**, 2420–2456.
- Ghosh, A., and Basu, A. (2016). Robust Estimation in Generalized Linear Models : The Density Power Divergence Approach. *Test*, **25(2)**, 269–290.
- Ghosh, A., Jaenada, M. and Pardo, L. (2020). Robust adaptive variable selection in ultra-high dimensional regression models based on the density power divergence loss. ArXiv preprint. arXiv:2004.05470
- Ghosh, A. and Majumdar, S. (2020). Ultrahigh-dimensional Robust and Efficient Sparse Regression using Non-Concave Penalized Density Power Divergence. *IEEE Trans. Info. Theory*, **66(12)**, 7812–7827.
- Ghosh A, Ponzi E, Sandanger T, and Thoresen M. (2021+) Robust Sure Independence Screening for Non-polynomial dimensional Generalized Linear Models. *Scand. J. Stat*, to appear.
- Ghosh, A. and Thoresen, M. (2021). A Robust Variable Screening procedure for Ultra-high dimensional data. *Stat. Meth. Med. Res.*, **30(8)**, 1816–1832.
- Gnanadesikan, R. and Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 81–124.

References

- Huang, J., Ma, S., and Zhang, C. H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Stat. Sinica*, 1603–1618.
- 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.
- Lambert-Lacroix, S., and Zwald, L. (2016). The adaptive BerHu penalty in robust regression. *J. Nonpar. Stat.*, **28(3)**, 487–514.
- Lambert-Lacroix, S., and Zwald, L. (2011). Robust regression through the Huber's criterion and adaptive lasso penalty. *Electron. J. Stat.*, **5**, 1015–1053.
- Li, G., Peng, H., Zhang, J. and Zhu, L. (2012a) Robust rank correlation based screening. *Ann Stat*, **40(3)**, 1846-1877.
- Li, R., Zhong, W. and Zhu, L. (2012b) Feature screening via distance correlation learning. *J Amer Statist Assoc*, **107**(499), 1129-1139.
- Meinshausen, N. and Buhlmann, P. (2010) Stability selection. *J Royal Stat Soc B*, **72(4)**, 417–473.
- Mu, W. and Xiong, S. (2014) Some notes on robust sure independence screening. *J App Stat*, **41(10)**, 2092–2102.
- Smucler, E., and Yohai, V. J. (2017). Robust and sparse estimators for linear regression models. *Comput. Stat. Data Anal.*, **111**, 116-130.
- Shah, R.D. and Samworth, R. J. (2013) Variable selection with error control: another look at stability selection. *J Royal Stat Soc B*, **75(1)**, 55–80.
- Szekely, G.J., Rizzo, M.L., and Bakirov, N.K. (2007). Measuring and testing dependence by correlation of distances. *Ann Stat*, **35**, 2769–2794.
- Wang, T., Zheng, L., Li, Z., Liu, H. (2017). A robust variable screening method for high-dimensional data. *J App Stat*, **44(10)**, 1839–1855.
- Zheng, Q., Gallagher, C., and Kulasekera, K. B. (2017). Robust adaptive Lasso for variable selection. *Commun. Stat. Theory Meth.*, **46(9)**, 4642–4659.
- Zou, H. (2006) The adaptive lasso and its oracle properties. *J Amer Stat Assoc*, **101**, 1418–1429.

THANK YOU!

Contact me at:
abhik.ghosh@isical.ac.in