Robust High-dimensional Statistics II: Minimum Divergence Approach

Dr. Abhik Ghosh

Indian Statistical Institute, Kolkata, India.

2022

Table of contents

Outline

Background: Minimum Distance Approach

- **Statistical Divergence Measure** $d(\cdot, \cdot)$ is a non-negative functional of two density (or distribution) functions that equals zero if and only if two arguments are identically equal.
- Examples: Kullback-Leibler (KL) Divergence, (Squared) Hellinger distance, L_2-Divergence

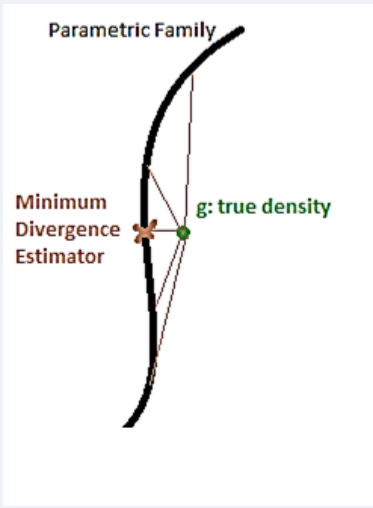
Minimum Divergence estimator

- Parametric inference: Model density family $\mathcal{F} = \{f_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$.
- Sample is observed from a population having true density g.
- **Minimum divergence estimator**:

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} d(\widehat{g}, f_\theta),$$

where $\widehat{g}$ is a non-parametric sample-based estimate of g.

- Motivation: MLE is the minimum KL divergence estimator!



Density based divergences has become extremely popular in robust statistical inference!
(Pardo, 2006; Basu et al., 2011)

Penalized Minimum Divergence estimators

$$\widehat{\theta} = \arg\min_{\theta} \left\{ d(\widehat{g}, f_\theta) + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \right\}$$

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

- If $\widehat{g}$ involves kernel, it is extremely difficult to obtain and study the resulting estimator in high-dimensional settings!
- We must use an appropriate distance for which g can be estimated without kernel (e.g., use the fact $gdx = dG(x)$)
- For useful distances, we can write $d(g, f_\theta) = \int \rho(\boldsymbol{z}, \theta) g(\boldsymbol{z}) d\boldsymbol{z}$+a term independent of θ. Then the corresponding empirical loss-function has the form:

$$\mathbf{d}(\widehat{\mathbf{g}}, \mathbf{f}_\theta) = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i}=\mathbf{1}}^{\mathbf{n}} \rho(\boldsymbol{z}_\mathbf{i}, \theta) + \mathbf{K}(\text{can be ignored}).$$

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. $f_{\boldsymbol{\beta}}(Y|\mathbf{x}) \equiv N(\mathbf{x}^T\boldsymbol{\beta}, \sigma^2)$

Minimum L_2 divergence Estimator

$$\begin{aligned}
\text{Minimize} \quad d_1(g(Y|\mathbf{x}), f_{\boldsymbol{\beta}}(Y|\mathbf{x})) &= \int \left[g(Y|\mathbf{x}) - f_{\boldsymbol{\beta}}(Y|\mathbf{x})\right]^2 dY \\
&= \int f_{\boldsymbol{\beta}}(Y|\mathbf{x})^2 dY - 2\int f_{\boldsymbol{\beta}}(Y|\mathbf{x}) g(Y|\mathbf{x}) dY + \int g(Y|\mathbf{x}) dY \\
&= \frac{1}{\sqrt{2\pi}\sigma} - \frac{2}{\sqrt{2\pi}\sigma}\int e^{-\frac{1}{2\sigma^2}(Y - \mathbf{x}^T\boldsymbol{\beta})^2} dG(Y|\mathbf{x}) + \text{constant}.
\end{aligned}$$

Minimum Distance (MD) Loss Function (Lozano et al., 2016)

$$L_n(\boldsymbol{\beta}) = -c \log\left[\sum_{i=1}^{n} e^{-\frac{1}{2c}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}\right], \quad \textbf{c is a scaling parameter.}$$

The MD-LASSO (Lozano et al., 2016)

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ -c \log\left[\sum_{i=1}^{n} e^{-\frac{1}{2c}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}\right] + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

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. $f_{\boldsymbol{\beta}}(Y|\mathbf{x}) \equiv N(\mathbf{x}^T\boldsymbol{\beta}, \sigma^2)$

Definition

Consider any sample of n points $(Y, \mathbf{x})$ and all possible corrupted samples $(Y', \mathbf{x}')$ that are obtained by replacing m of the original points by arbitrary values. The breakdown point of a regression estimator $\widehat{\boldsymbol{\beta}}$ at the sample $(Y, \mathbf{x})$ is defined as

$$\epsilon_n^*(\widehat{\boldsymbol{\beta}}; Y, \mathbf{x}) = \min\left\{\frac{m}{n} : \sup_{(Y', \mathbf{x}')} ||\widehat{\boldsymbol{\beta}}(Y', \mathbf{x}')||_2 = \infty\right\}.$$

Asymptotic breakdown point: $\epsilon^* = \lim_{n\to\infty} \epsilon_n^*(\widehat{\boldsymbol{\beta}}; Y, \mathbf{x})$.

Result (Lozano et al., 2016)

For any finite choice of c, let $Q_c(\boldsymbol{\beta})$ denote the MD-Lasso objective function and consider the non-empty set $B_c = \{\boldsymbol{\beta} : \boldsymbol{\beta}$ is a local optimum for the MD-Lasso problem and $Q_c(\boldsymbol{\beta}) \leq Q_c(0)\}$.
For every $\alpha \in (0, 1)$, the breakdown point of any solution in B_c is at least α.
The MD-Lasso can tolerate at least αn arbitrarily corrupted observations and still produce estimates having bounded ℓ_2-norm!

Lozano et al. (2016) derived the consistency of the MD-LASSO under the high-dimensional set-up assuming Restricted Strong Convexity in a local neighborhood of true parameter value, when

- MD-LASSO belongs to the same local convexity neighborhood!
- MD-LASSO is computed under additional restriction $||\beta||_2 \leq r$ (feasibility)!

Outline

Data: $X_1, \ldots, X_n \sim g$ (IID), Model: $f_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$;

Density Power Divergence is a generalization of the KL-divergence (Basu et al., 1998)

$$d_\alpha(g, f_{\boldsymbol{\theta}}) = \int f_{\boldsymbol{\theta}}^{1+\alpha} - \frac{1+\alpha}{\alpha} \int f_{\boldsymbol{\theta}}^{\alpha} g + \frac{1}{\alpha} \int g^{1+\alpha}, \quad \text{if} \quad \alpha > 0,$$

$$d_0(g, f_{\boldsymbol{\theta}}) = \lim_{\alpha \to 0} d_\alpha(g, f_{\boldsymbol{\theta}}) = \int g \log(g/f_{\boldsymbol{\theta}}) = \text{KL-divergence}.$$

The Minimum Density Power Divergence Estimation (MDPDE)

The MDPDE of $\boldsymbol{\theta}$ is the minimizer of $d_\alpha(\widehat{g}, f_{\boldsymbol{\theta}})$ with respect to $\boldsymbol{\theta}$, or equivalently the minimizer of

$$\mathbf{H_n(\theta)} = \int \mathbf{f_\theta^{1+\alpha}} - \frac{\mathbf{1+\alpha}}{\alpha} \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{f_\theta^\alpha(X_i)}.$$

Properties of the MDPDE under IID data (Basu et al., 2011)

- MDPDE at $\alpha = 0$ is the MLE, most efficient but highly non-robust (unbounded IF).
- Consistent and Asymptotically normal for all $\alpha \geq 0$.
- Large α = more robust, less efficient. Small α = less robust, more efficient.
- The loss in efficiency is not quite significant at small $\alpha > 0$ in most cases.

Independent Non-homogeneous (INH) Set-up

- Data: $X_i \sim g_i$ independently for $i = 1, \ldots, n$.
- Model: $X_i \sim f_{i,\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ for $i = 1, \ldots, n$.
- Example: Fixed-design regressions.

The MDPDE under INH Set-up (Ghosh and Basu, 2013)

- The MDPDE of $\boldsymbol{\theta}$ is defined as the minimizer of average DPD measure $\frac{1}{n}\sum_{i=1}^{n} d_\alpha(\widehat{g}_i, f_{i,\boldsymbol{\theta}})$.
- This leads to the objective function

$$\mathbf{H_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left[\int f_{i,\theta}^{1+\alpha} - \frac{1+\alpha}{\alpha} f_{i,\boldsymbol{\theta}}^{\alpha}(X_i)\right].}$$

Properties are similar to the IID case (under slightly different assumptions).

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.

For each i, $y_i \sim f_{i,\boldsymbol{\theta}}$, where $f_{i,\boldsymbol{\theta}} \equiv N(\mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2)$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$. So, it belongs to the INH Set-up!!

The DPD-based loss function (Ghosh and Basu, 2013)

- According to the INH set-up, the MDPDE should be the minimizer of

$$H_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left[\int f_{i,\boldsymbol{\theta}}^{1+\alpha} - \frac{1+\alpha}{\alpha} f_{i,\boldsymbol{\theta}}^{\alpha}(X_i)\right].$$

- For Normal error density, it simplifies to

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

Durio and Isaia (2011) considered it from random-design perspective: same objective function!

As $\alpha \downarrow 0$, $L_n^{\alpha}(\boldsymbol{\beta}, \sigma)$ coincides (in a limiting sense) with the negative log-likelihood (plus one).

(why? think L-Hospital's rule.)

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 DPD

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

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

Minimum Penalized DPD estimator (MPDPD) of (β, σ) at any given $\alpha \in [0, 1]$

Minimize the penalized DPD objective function $Q_n^{\alpha}(\beta, \sigma)$ with respect to both β and σ!

- As $\alpha \downarrow 0$, this becomes non-concave penalized negative log-likelihood (non-robust).
- At $\alpha = 1$, it has a link to MD-LASSO if the penalty is ℓ_1 (but not exactly coincide - why?)

- Zang et al. (2017): Empirically studied the MPDPDE with grouped lasso penalty (without theory).
- Ghosh and Majumdar (2020): Complete theory, including IF, for general non-concave penalties!

Non-convex minimization problem:
Minimize the penalized DPD objective function $Q_n^\alpha(\boldsymbol{\beta}, \sigma)$ with respect to both $\boldsymbol{\beta}$ and σ!

Main idea

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

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

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

Solving Individual Minimization Problems

- Update $\boldsymbol{\beta}$ using a Concave-Convex Procedure (Yuille and Rangarajan, 2003):

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

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

- Update σ using gradient descent.

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

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

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

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

$$\text{RobHBIC}(\lambda) = \log(\widehat{\sigma}^{\mathbf{2}}) + \frac{\log\log(\mathbf{n})\log \mathbf{p}}{\mathbf{n}} \|\widehat{\boldsymbol{\beta}}\|_{\mathbf{0}}, \tag{1}$$

where $\widehat{\sigma}^2$ is estimated each time as the MPDPDE of σ^2.

Select the optimal λ^* that minimizes the HBIC over a pre-determined set of values Λ_n:

$$\lambda^* = \underset{\lambda \in \Lambda_n}{\text{argmin}}\, \text{RobHBIC}(\lambda).$$

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.

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

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

$$\mathbf{L_n^{(\alpha)}}(\boldsymbol{\beta}, \sigma) = \frac{\mathbf{1}}{\sigma^{\alpha}} \mathbf{M_f^{(\alpha)}} - \frac{\mathbf{1}+\alpha}{\alpha} \frac{\mathbf{1}}{\mathbf{n}\sigma^{\alpha}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{f}^{\alpha}\left(\frac{\mathbf{y_i} - \boldsymbol{x}_\mathbf{i}^\mathbf{t}\boldsymbol{\beta}}{\sigma}\right) + \frac{\mathbf{1}}{\alpha}. \tag{2}$$

The General MPDPDE

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

Outline

MPDPDE Functional

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

where

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

MPDPDE Functional

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

where

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

$$\psi_\alpha((y,\boldsymbol{x});\boldsymbol{\theta}) = \nabla L_\alpha^*((y,\boldsymbol{x});\boldsymbol{\theta}) = \frac{(1+\alpha)}{\sigma^{\alpha+1}} \begin{bmatrix} \psi_{1,\alpha}\left(\frac{y-\boldsymbol{x}^T\boldsymbol{\beta}}{\sigma}\right)\boldsymbol{x} \\ \psi_{2,\alpha}\left(\frac{y-\boldsymbol{x}^T\boldsymbol{\beta}}{\sigma}\right) \end{bmatrix}, \tag{3}$$

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

$$\begin{aligned} \psi_{1,\alpha}(s) &= u(s) f^\alpha(s), \\ \psi_{2,\alpha}(s) &= \{su(s)+1\} f^\alpha(s) - \frac{\alpha}{\alpha+1} M_f^{(\alpha)}. \end{aligned} \tag{4}$$

Assumptions

Suppose that the **penalty $\widetilde{p}_\lambda(s) = p_\lambda(|s|)$ is twice differentiable in s** and assume the following:

- $M_f^{(\alpha)}$, $\int \psi_\alpha((y, \boldsymbol{x}); \boldsymbol{\theta}) dG(y, \boldsymbol{x})$, $\boldsymbol{J}_\alpha(G; \boldsymbol{\theta}) = \int \nabla \psi_\alpha((y, \boldsymbol{x}); \boldsymbol{\theta}) dG(y, \boldsymbol{x})$ are all finite.
- $\boldsymbol{J}^*_\alpha(G; \boldsymbol{\theta}) := \left[\boldsymbol{J}_\alpha(G; \boldsymbol{\theta}) + \text{diag}\left\{\widetilde{\boldsymbol{P}}^{**}_\lambda(\boldsymbol{\beta}), 0\right\}\right]$ is invertible at $\boldsymbol{\theta} = \boldsymbol{\theta}_g = (\boldsymbol{\beta}_g, \sigma_g)^T = \boldsymbol{T}_\alpha(G)$.

Define: $\widetilde{\boldsymbol{P}}^{**}_\lambda(\boldsymbol{\beta}) = \text{diag}\left\{\widetilde{p}''_\lambda(\beta_1), \ldots, \widetilde{p}''_\lambda(\beta_p)\right\}$.

Main Result (Ghosh and Majumdar, 2020)

Whenever it exists, the influence function of the MPDPDE functional $\boldsymbol{T}_\alpha$ at G is given by

$$\mathcal{IF}((\mathbf{y_t}, \boldsymbol{x_t}), \boldsymbol{T}_\alpha, \mathbf{G}) = -\boldsymbol{J}^*_\alpha(\mathbf{G}; \boldsymbol{\theta}^{\mathbf{g}})^{-1} \begin{bmatrix} \frac{(1+\alpha)}{\sigma_g^{\alpha+1}} \psi_{1,\alpha}\left(\frac{y_t - \boldsymbol{x}_t^T \boldsymbol{\beta}_g}{\sigma_g}\right) \boldsymbol{x}_t + \widetilde{\boldsymbol{P}}^*_\lambda(\boldsymbol{\beta}_g) \\ \frac{(1+\alpha)}{\sigma_g^{\alpha+1}} \psi_{2,\alpha}\left(\frac{y_t - \boldsymbol{x}_t^T \boldsymbol{\beta}_g}{\sigma_g}\right) \end{bmatrix}. \tag{5}$$

Further, if Θ is compact, then the above IF exists for all $(y_t, \boldsymbol{x}_t)$.

Assumptions

Consider the general penalty $p_\lambda(|s|)$ where $p_\lambda(s)$ **is twice differentiable in** s and assume that true value $\boldsymbol{\beta}_g = (\boldsymbol{\beta}_{1g}^T, \mathbf{0}_{p-s}^T)^T$, where $\boldsymbol{\beta}_{1g}$ contains all and only s-non-zero elements of $\boldsymbol{\beta}_g$.

Define: $\boldsymbol{P}_\lambda^*(\boldsymbol{v}) = (p'_\lambda(|v_1|)sign(v_1), \ldots, p'_\lambda(|v_q|)sign(v_q))$, and $\boldsymbol{P}_\lambda^{**}(\boldsymbol{v}) = \text{diag}\{p''_\lambda(|v_1|), \ldots, p''_\lambda(|v_q|)\}$, for any $\boldsymbol{v} = (v_1, \ldots, v_q)^T$ with all non-zero elements.

Main Result (Ghosh and Majumdar, 2020)

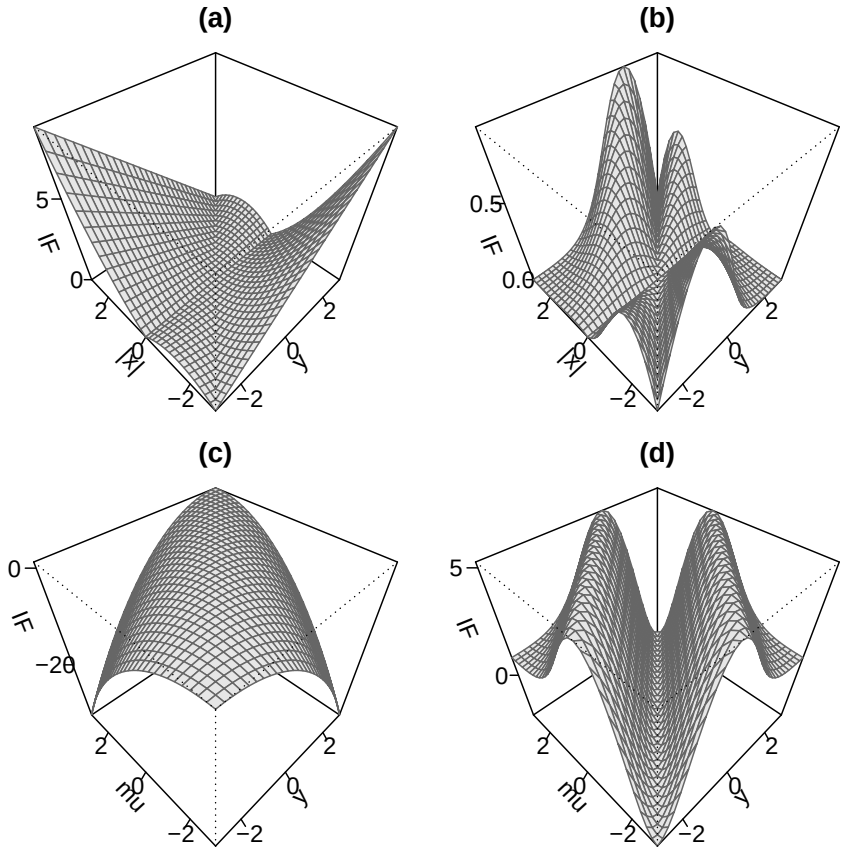
Let us denote $\boldsymbol{\theta}_g = (\boldsymbol{\beta}_{1g}^T, \mathbf{0}_{p-s}^T, \sigma_g)^T = \boldsymbol{T}_\alpha(G) = (\boldsymbol{T}_{1,\alpha}^\beta(G)^T, \boldsymbol{T}_{2,\alpha}^\beta(G)^T, T_\alpha^\sigma(G))^T$. Then, whenever the associated quantities exists, the influence function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

$$\mathcal{IF}((\mathbf{y_t}, \boldsymbol{x_t}), (\boldsymbol{T}_{1,\alpha}^\beta, \mathbf{T}_\alpha^\sigma), \mathbf{G}) = -\mathbf{S}_\alpha(\mathbf{G}; \boldsymbol{\theta_g})^{-1} \begin{bmatrix} \frac{(1+\alpha)}{(\sigma_g)^{\alpha+1}} \psi_{1,\alpha}\left(\frac{y_t - \boldsymbol{x}_t^T\boldsymbol{\beta}_g}{\sigma_g}\right)\boldsymbol{x}_{1,t} + \boldsymbol{P}_\lambda^*(\boldsymbol{\beta}_{1g}) \\ \frac{(1+\alpha)}{(\sigma_g)^{\alpha+1}} \psi_{2,\alpha}\left(\frac{y_t - \boldsymbol{x}_t^T\boldsymbol{\beta}_g}{\sigma_g}\right) \end{bmatrix},$$

$$\mathbf{S}_\alpha(G;\boldsymbol{\theta}) = -\frac{(1+\alpha)}{\sigma^{\alpha+2}} E_G \begin{bmatrix} J_{11,\alpha}\left(\frac{y-\boldsymbol{x}^T\boldsymbol{\beta}}{\sigma}\right)\boldsymbol{x}_1\boldsymbol{x}_1^T & J_{12,\alpha}\left(\frac{y-\boldsymbol{x}^T\boldsymbol{\beta}}{\sigma}\right)\boldsymbol{x}_1 \\ J_{12,\alpha}\left(\frac{y-\boldsymbol{x}^T\boldsymbol{\beta}}{\sigma}\right)\boldsymbol{x}_1^T & J_{22,\alpha}\left(\frac{y-\boldsymbol{x}^T\boldsymbol{\beta}}{\sigma}\right) \end{bmatrix} + \begin{bmatrix} \boldsymbol{P}_\lambda^{**}(\boldsymbol{\beta}_1) & \mathbf{0}_s \\ \mathbf{0}_s^T & 0 \end{bmatrix}.$$

Further, if Θ is compact, then the IF exists in both cases for all $(y_t, \boldsymbol{x}_t)$.

IF for MPDPDE: Example for normal error



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, $(y_t, \mathbf{x}_t^T\boldsymbol{\beta})$ 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

Theorem (Ghosh and Majumdar, 2020)

Consider the general penalized DPD loss function $Q^{\alpha}_{n,\lambda}(\boldsymbol{\theta})$ for a fixed $\alpha \geq 0$ and a general error density f and a penalty function p_λ satisfying Assumption (P). Then, $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}, \widehat{\sigma})$ is a strict local minimizer of $Q^{\alpha}_{n,\lambda}(\boldsymbol{\theta})$ if and only if

$$\frac{1+\alpha}{n\widehat{\sigma}^{\alpha+1}} \sum_{i=1}^{n} \psi_{1,\alpha}(r_i(\widehat{\boldsymbol{\theta}}))\boldsymbol{x}_{1i} + \boldsymbol{P}^{*}_{\lambda}(\widehat{\boldsymbol{\beta}}_1) = \boldsymbol{0}, \tag{6}$$

$$\left\| \frac{1+\alpha}{n\lambda\widehat{\sigma}^{\alpha+1}} \sum_{i=1}^{n} \psi_{1,\alpha}(r_i(\widehat{\boldsymbol{\theta}}))\boldsymbol{x}_{2i} \right\|_{\infty} < \rho(p_\lambda) = p'_\lambda(0+)/\lambda, \tag{7}$$

$$\frac{1+\alpha}{n\widehat{\sigma}^{\alpha+1}} \sum_{i=1}^{n} \psi_{2,\alpha}(r_i(\widehat{\boldsymbol{\theta}})) = 0, \tag{8}$$

$$\Lambda_{\min}\left(-\frac{1+\alpha}{n\widehat{\sigma}^{\alpha+2}} \sum_{i=1}^{n} \begin{bmatrix} J_{11,\alpha}(r_i(\widehat{\boldsymbol{\theta}}))\boldsymbol{x}_{1i}\boldsymbol{x}_{1i}^T & J_{12,\alpha}(r_i(\widehat{\boldsymbol{\theta}}))\boldsymbol{x}_{1i} \\ J_{12,\alpha}(r_i(\widehat{\boldsymbol{\theta}}))\boldsymbol{x}_{1i}^T & J_{22,\alpha}(r_i(\widehat{\boldsymbol{\theta}})) \end{bmatrix} \right) > \zeta(p_\lambda, \widehat{\boldsymbol{\beta}}_1) \tag{9}$$

$$= \lim_{\epsilon \downarrow 0} \max_{1 \leq j \leq s} \sup_{t_1 < t_2 \in (|\beta_j|-\epsilon, |\beta_j|+\epsilon)} -\frac{p'_\lambda(t_2) - p'_\lambda(t_1)}{t_2 - t_1}.$$

where $\widehat{\boldsymbol{\beta}}_1$ contains non-zero components of $\widehat{\boldsymbol{\beta}}$, and $\boldsymbol{x}_i = (\boldsymbol{x}_{1i}^T, \boldsymbol{x}_{2i}^T)^T$ is the corresponding partition of $\boldsymbol{x}_i$ for each i with $\boldsymbol{x}_{1i}$ having the same dimension as $\widehat{\boldsymbol{\beta}}_1$.

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** [Ghosh and Majumdar, 2020]

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

for $c > 0$, and

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

Results: Weak Oracle Properties

(A1) $||\boldsymbol{x}^{(j)}|| = O(\sqrt{n})$ for $j = 1, \ldots, p$, where $\boldsymbol{x}^{(j)}$ is the j^{th} column of $\mathbf{X}$.

(A2) For some $C \in (0,1)$, $\tau_1 \in [0, 0.5]$, we have

$$\left\| \left(\mathbf{X}_S^{*T} \nabla^2 L_n^{\alpha}(\boldsymbol{\theta}_0) \mathbf{X}_S^{*} \right)^{-1} \right\|_{\infty} = O\left(\frac{b_s}{n} \right), \quad \left\| \left(\mathbf{X}_N^{*T} \nabla^2 L_n^{\alpha}(\boldsymbol{\theta}_0) \mathbf{X}_S^{*} \right) \left(\mathbf{X}_S^{*T} \nabla^2 L_n^{\alpha} \boldsymbol{\theta}_0) \mathbf{X}_S^{*} \right)^{-1} \right\|_{\infty} < \min \left\{ \frac{C p'_{\lambda}(0+)}{p'_{\lambda}(d_n)}, O(n^{\tau_1}) \right\},$$

$$\max_{(\boldsymbol{\delta},\sigma) \in \mathcal{N}_0} \max_{1 \le j \le p+1} \left\{ \Lambda_{\max} \left(\mathbf{X}_S^{*} \left[\sum_{i=1}^{n} \nabla^2_{(\boldsymbol{\delta},\sigma)} \psi_{\alpha,j} \left((y_i, \boldsymbol{x}_i), (\boldsymbol{\delta}^T, \mathbf{0}_{p-s}^T, \sigma)^T \right) \right] \mathbf{X}_S^{*T} \right) \right\} = O(n).$$

(A3) Let $s = O(n^{\tau_0})$. For some $\tau \in (0, 0.5]$, define $\tau^* = \min\{0.5, 2\tau - \tau_0\} - \tau_1$. Then,

$$d_n \ge \log n / n^{\tau}, \quad b_s = o(\min\{n^{1/2-\tau} \sqrt{\log n}, n^{\tau} / s \log n\}), \quad p'_{\lambda}(d_n) = o\left(\frac{\log n}{b_s n^{\tau}} \right), \quad \lambda \ge \frac{(\log n)^2}{n^{\tau^*}}.$$

Also, $\max_{1 \le j \le p} ||\boldsymbol{x}^{(j)}||_{\infty} = o\left(n^{\tau^*} / \sqrt{\log n} \right)$ and $\max_{(\boldsymbol{\delta},\sigma) \in \mathcal{N}_0} \zeta(p_{\lambda}; \boldsymbol{\delta}) = o\left(\max_{(\boldsymbol{\delta},\sigma) \in \mathcal{N}_0} \Lambda_{\min} \left[\frac{1}{n} \mathbf{X}_S^{*T} \nabla^2 L_n^{\alpha}(\boldsymbol{\delta}, \sigma) \mathbf{X}_S^{*} \right] \right)$.

(A4) For any $\boldsymbol{a} \in \mathbb{R}^n$ and $0 < \epsilon < \|\boldsymbol{a}\| / \|\boldsymbol{a}\|_{\infty}$, there is $c_1 > 0$ such that

$$P\left(\left| \frac{1+\alpha}{\sigma_0^{\alpha+1}} \sum_{i=1}^{n} a_i \psi_{1,\alpha}(r_i(\boldsymbol{\theta}_0)) \right| > \|\boldsymbol{a}\| \epsilon \right) \le 2e^{-c_1 \epsilon^2}.$$

Theorem (Ghosh and Majumdar, 2020)

Let $s = o(n)$, $\log p = O(n^{1-2\tau^*})$ and Assumptions (P), (A1)–(A4) hold for the given α. Then, there exist MNPDPDEs $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_S^T, \widehat{\boldsymbol{\beta}}_N^T)^T$ of $\boldsymbol{\beta}$, with $\widehat{\boldsymbol{\beta}}_S \in \mathbb{R}^s$, and $\widehat{\sigma}$ of σ such that $(\widehat{\boldsymbol{\beta}}, \widehat{\sigma})$ is a (strict) local minimizer of $Q_{n,\lambda}^{\alpha}(\boldsymbol{\theta})$, with

$$\widehat{\boldsymbol{\beta}}_{\mathbf{N}} = \mathbf{0}_{\mathbf{p}-\mathbf{s}}, \quad \left\| \widehat{\boldsymbol{\beta}}_{\mathbf{S}} - \boldsymbol{\beta}_{\mathbf{S0}} \right\|_{\infty} = \mathbf{O}\left(\frac{\log \mathbf{n}}{\mathbf{n}^{\tau}} \right), \quad |\widehat{\sigma} - \sigma_{\mathbf{0}}| = \mathbf{O}\left(\frac{\log \mathbf{n}}{\mathbf{n}^{\tau}} \right),$$

holding with probability $\ge 1 - (2/n)(1 + s + (p-s)\exp[-n^{1-2\tau^*}])$.

Stronger Assumptions

(A2*) For some $c > 0$, the design matrix $\mathbf{X}$ satisfies

$$\min_{(\boldsymbol{\delta},\sigma)\in\mathcal{N}_0} \Lambda_{\min}\left[\mathbf{X}_S^{*T}\nabla^2 L_n^{\alpha}(\boldsymbol{\delta},\sigma)\mathbf{X}_S^{*}\right] \geq cn, \quad \left\|\left(\mathbf{X}_N^{*T}\nabla^2 L_n^{\alpha}(\boldsymbol{\theta}_0)\mathbf{X}_S^{*}\right)\right\|_{2,\infty} = O(n),$$

$$\max_{(\boldsymbol{\delta},\sigma)\in\mathcal{N}_0}\max_{1\leq j\leq p+1}\Lambda_{\max}\left(\mathbf{X}_S^{*}\left[\sum_{i=1}^{n}\nabla^2_{(\boldsymbol{\delta},\sigma)}\psi_{\alpha,j}\left((y_i,\boldsymbol{x}_i),(\boldsymbol{\delta}^T,\mathbf{0}_{p-s}^T,\sigma)^T\right)\right]\mathbf{X}_S^{*T}\right) = O(n),$$

$$E\left\|\frac{1+\alpha}{\sigma_0^{\alpha+1}}\sum_{i=1}^{n}\psi_{1,\alpha}(r_i(\boldsymbol{\theta}_0))\boldsymbol{x}_{Si}\right\|_2^2 = O(sn), \quad E\left|\frac{1+\alpha}{\sigma_0^{\alpha+1}}\sum_{i=1}^{n}\psi_{2,\alpha}(r_i(\boldsymbol{\theta}_0))\right|^2 = O(n),$$

(A3*) For some $\tau \in (0, 0.5]$, we have

$$p'_{\lambda}(d_n) = O(n^{-1/2}), \quad d_n \gg \lambda \gg \min\left\{s^{1/2}n^{-1/2}, n^{\frac{\tau-1}{2}}\sqrt{\log n}\right\},$$

$$\max_{1\leq j\leq p}\|\boldsymbol{x}^{(j)}\|_{\infty} = o(n^{(1-\tau)/2}/\sqrt{\log n}), \quad \max_{(\boldsymbol{\delta},\sigma)\in\mathcal{N}_0}\zeta(p_{\lambda};\boldsymbol{\delta}) = o(1).$$

Theorem (Ghosh and Majumdar, 2020)

Suppose $s \ll n$ and $\log p = O(n^{\tau^*})$ for some $\tau^* \in (0, 0.5)$ and Assumptions (P), (A1), (A2), (A3*) and (A4) hold at a fixed $\alpha \geq 0$. Then, there exists a strict minimizer (strict MPDPDE) $(\widehat{\boldsymbol{\beta}}^T, \widehat{\sigma})^T = \left((\widehat{\boldsymbol{\beta}}_S^T, \widehat{\boldsymbol{\beta}}_N^T)^T, \widehat{\sigma}\right)$, with $\widehat{\boldsymbol{\beta}}_S \in \mathbb{R}^s$, that satisfies the following with probability tending to 1 as $n \to \infty$.*

$$\widehat{\boldsymbol{\beta}}_{\mathbf{N}} = \mathbf{0}, \quad \left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\mathbf{0}}\right\| = \mathbf{O}(\sqrt{\mathbf{s}/\mathbf{n}}), \quad |\widehat{\sigma} - \sigma_{\mathbf{0}}| = \mathbf{O}(\mathbf{n}^{-\mathbf{1/2}}).$$

Additional Assumption

(A5) The penalty, loss function and design matrix satisfy the following conditions:

$$p'_\lambda(d_n) = O((sn)^{-\frac{1}{2}}), \qquad \max_{1\le i\le n} E\left|\psi_{k,\alpha}(r_i(\boldsymbol{\theta}_0))\right|^3 = O(1), k = 1, 2,$$

$$\min_{(\boldsymbol{\delta},\sigma)\in\mathcal{N}_0} \Lambda_{\min}\left[\mathbf{X}_S^{*T}\boldsymbol{\Sigma}_{\alpha *}((\boldsymbol{\delta}^T, \mathbf{0}_{p-s}, \sigma)^T)\mathbf{X}_S^*\right] \ge cn, \quad \sum_{i=1}^{n}\left[\boldsymbol{x}_{Si}^{*T}\left(\mathbf{X}_S^{*T}\boldsymbol{\Sigma}^*(\boldsymbol{\theta}_0)\mathbf{X}_S^*\right)^{-1}\boldsymbol{x}_{Si}^*\right]^{3/2} = o(1).$$

where

$$\boldsymbol{\Sigma}_\alpha^*(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{K}_{11}^{(\alpha)}(\boldsymbol{\theta}) & \boldsymbol{K}_{12}^{(\alpha)}(\boldsymbol{\theta}) \\ \boldsymbol{K}_{12}^{(\alpha)}(\boldsymbol{\theta}) & \boldsymbol{K}_{22}^{(\alpha)}(\boldsymbol{\theta}) \end{pmatrix}, \quad \boldsymbol{K}_{ij,\alpha}(\boldsymbol{\theta}) = \frac{(1+\alpha)^2}{\sigma^{2\alpha+2}}\,\mathrm{diag}\left\{\psi_{i,\alpha}(r_1(\boldsymbol{\theta}))\psi_{j,\alpha}(r_1(\boldsymbol{\theta})), \ldots, \psi_{i,\alpha}(r_n(\boldsymbol{\theta}))\psi_{j,\alpha}(r_n(\boldsymbol{\theta}))\right\}.$$

Theorem (Ghosh and Majumdar, 2020)

In addition to the assumptions of te previous theorem, suppose that $s = o(n^{1/3})$ and (A5) holds. Then, the strict MNPDPDE $(\widehat{\boldsymbol{\beta}}, \widehat{\sigma})$ satisfies the following results with probability tending to 1 as $n \to \infty$.

1. $\widehat{\boldsymbol{\beta}}_N = \mathbf{0}$, *where* $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_S^T, \widehat{\boldsymbol{\beta}}_N^T)^T$ *with* $\widehat{\boldsymbol{\beta}}_S \in \mathbb{R}^s$.
2. *Let* $\boldsymbol{A}_n \in \mathbb{R}^{q\times(s+1)}$ *such that* $\boldsymbol{A}_n\boldsymbol{A}_n^T \to \boldsymbol{G}$ *as* $n \to \infty$, *where* $\boldsymbol{G}$ *is symmetric and positive definite. Then,*

$$\boldsymbol{A_n}\left(\mathbf{X_S^{*T}}\boldsymbol{\Sigma}_\alpha^*(\boldsymbol{\theta_0})\mathbf{X_S^*}\right)^{-1/2}\left(\mathbf{X_S^{*T}}\boldsymbol{\Sigma}_\alpha(\boldsymbol{\theta_0})\mathbf{X_S^*}\right)\left((\widehat{\boldsymbol{\beta}}_\mathbf{S}, \widehat{\sigma}) - (\boldsymbol{\beta}_{\mathbf{S0}}, \sigma_\mathbf{0})\right) \xrightarrow{\mathcal{D}} \mathbf{N_q}(\mathbf{0_q}, \boldsymbol{G}). \tag{10}$$

Outline

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

$$f(y; \boldsymbol{x}^T\beta) = \exp\{y\theta - b(\theta) + c(y)\}, \quad \text{with} \quad E[Y|\boldsymbol{x}] = b'(\theta) = g^{-1}(\boldsymbol{x}^T\beta), \tag{11}$$

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)

$$\mathbf{L_n^{(\alpha)}}(\beta) = \frac{\mathbf{1}}{\mathbf{n(1+\alpha)}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \left[\int \mathbf{f(y, x_i^t\beta)^{\alpha+1} dy} - \frac{\mathbf{1+\alpha}}{\alpha} \mathbf{f(y_i, x_i^t\beta)^{\alpha}} + \frac{\mathbf{1}}{\alpha} \right].$$

Minimum Penalized DPD Estimator (MPDPDE)

$$\widehat{\beta} = \arg\min_{\beta} \left\{ L_n^{(\alpha)}(\beta, \sigma) + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \right\}.$$

- **(A1)** The l_2-norm of each column of $\boldsymbol{X}$ is $O(\sqrt{n})$.

(A2) $\left\|\boldsymbol{\Psi}_{n,\alpha}(\boldsymbol{\beta}_0)^{-1}\right\|_\infty = O\left(\frac{b_s}{n}\right)$, $\left\|\widetilde{\boldsymbol{\Psi}}_{n,\alpha}(\boldsymbol{\beta}_0)\boldsymbol{\Psi}_{n,\alpha}(\boldsymbol{\beta}_0)^{-1}\right\|_\infty < \min\left\{\frac{Cp'_\lambda(0+)}{p'_\lambda(d_n)}, O(n^{\tau_1})\right\}$ and $\max_{\boldsymbol{\delta}\in\mathcal{N}_0}\max_{1\le j\le p+1}\left\{\Lambda_{\max}\left(\boldsymbol{X}_S\left[\nabla^2_{\boldsymbol{\delta}}\Gamma_{j,\alpha}(\boldsymbol{\delta})\right]\boldsymbol{X}_S^t\right)\right\} = O(n)$, for some constants $C\in(0,1)$, $\tau_1\in[0,0.5]$, where $\mathcal{N}_0 = \{\boldsymbol{\delta}\in\mathbb{R}^s : ||\boldsymbol{\delta}-\boldsymbol{\beta}_{S0}||_\infty \le d_n\}$, $\boldsymbol{\Gamma}_\alpha(\boldsymbol{\delta}) = \left(\Gamma_{1,\alpha}(\boldsymbol{\delta}),\ldots,\Gamma_{p,\alpha}(\boldsymbol{\delta})\right)^t = \sum_{i=1}^n \psi_\alpha\left(y_i;\boldsymbol{x}_{iS}^t\boldsymbol{\delta}\right)$. and $\nabla^2_{\boldsymbol{\delta}}$ denotes the second-order derivative with respect to $\boldsymbol{\delta}$.

(A3) For some $\tau\in(0,0.5]$, we have $d_n \ge \log n/n^\tau$ and $b_s = o(\min\{n^{1/2-\tau}\sqrt{\log n}, n^\tau/s\log n\})$. Further, with $s = O(n^{\tau_0})$, we define $\tau^* = \min\{0.5, 2\tau-\tau_0\}-\tau_1$. Then, we have $p'_\lambda(d_n) = o\left(\frac{\log n}{b_s n^\tau}\right)$, $\lambda \ge (\log n)^2 n^{-\tau^*}$, and $\max_{\boldsymbol{\delta}\in\mathcal{N}_0}\zeta(p_\lambda;\boldsymbol{\delta}) = o\left(\max_{\boldsymbol{\delta}\in\mathcal{N}_0}\Lambda_{\min}\left[n^{-1}\boldsymbol{\Psi}_{n,\alpha}((\boldsymbol{\delta}^t,\mathbf{0}_{p-s}^t))\right]\right)$, where ζ denote the local concavity of the penalty p_λ as defined in **?**, Definition 1. Also, the maximum (in absolute) element of the design matrix $\boldsymbol{X}$ is of order $o\left(n^{\tau^*}/\sqrt{\log n}\right)$.

(A4) For any $\boldsymbol{a}\in\mathbb{R}^n$ and $0<\epsilon<\|\boldsymbol{a}\|/\|\boldsymbol{a}\|_\infty$, there exists a $c_1>0$ such that

$$P\left(\left|\sum_{i=1}^n a_i\psi_\alpha(y_i,\boldsymbol{x}_i^t\boldsymbol{\beta}_0)\right| > \|\boldsymbol{a}\|\epsilon\right) \le 2e^{-c_1\epsilon^2}.$$

(A5) $p'_\lambda(d_n) = O((sn)^{-1/2})$, $\max_{1\le i\le n} E\left|\psi_\alpha(y_i,\boldsymbol{x}_i^t\boldsymbol{\beta}_0)\right|^3 = O(1)$,

Theorem (Ghosh, 2021+)

Under the set-up of ultra high-dimensional GLMs (11) with $s = o(n)$ and $\log p = O(n^{1-2\tau^*})$, let us assume that Assumptions (A1)–(A5) and (P) hold for some fixed $\alpha \geq 0$.

Then, there exist MNPDPDEs $\widehat{\beta} = (\widehat{\beta}_S^t, \widehat{\beta}_N^t)^t$ of β, with $\widehat{\beta}_S \in \mathbb{R}^s$, such that $\widehat{\beta}$ is a (strict) local minimizer (MPDPDE) and satisfies the following:

a) $\widehat{\beta}_N = \mathbf{0}_{p-s}$, *and* $\left\|\widehat{\beta}_S - \beta_{0S}\right\|_\infty = O\left(n^{-\tau}\log n\right)$,
with probability at least $P_n = 1 - (2/n)(1 + s + (p-s)\exp[-n^{1-2\tau^*}])$.

b) *Let* $\boldsymbol{A}_n \in R^{q\times(s+1)}$ *such that* $\boldsymbol{A}_n\boldsymbol{A}_n^t \to \boldsymbol{G}$ *as* $n \to \infty$, *where* $\boldsymbol{G}$ *is symmetric and positive definite. Then, with probability tending to one,*

$$\boldsymbol{A_n\Omega_{n,\alpha}(\beta_0)^{-1/2}\Psi_{n,\alpha}(\beta_0)}\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{S}} - \boldsymbol{\beta_{0S}}\right) \xrightarrow{\mathcal{D}} \boldsymbol{N_q(0_q, G)}.$$

$$\boldsymbol{\Psi}_{n,\alpha}(\beta) = \left(\boldsymbol{X}_S^t\boldsymbol{\Sigma}_{n,\alpha}(\beta_0)\boldsymbol{X}_S\right), \qquad \boldsymbol{\Omega}_{n,\alpha}(\beta) = \left(\boldsymbol{X}_S^t\boldsymbol{\Sigma}^*_{n,\alpha}(\beta_0)\boldsymbol{X}_S\right),$$

where $\boldsymbol{\Sigma}_{n,\alpha}(\beta) = \text{diag}\left\{-\frac{\partial}{\partial\eta}\psi_\alpha(y_i,\eta)\big|_{\eta=\boldsymbol{x}_i^t\beta} \;:\; i = 1,\ldots,n\right\}$
and $\boldsymbol{\Sigma}^*_{n,\alpha}(\beta) = \text{diag}\left\{\psi_\alpha(y_i, \boldsymbol{x}_i^t\beta)^2 \;:\; i = 1,\ldots,n\right\}$.

Theorem (Ghosh, 2021+)

Under the ultra high-dimensional GLMs (11) with $s \ll n$, $\log p = O(n^{\tau^*})$ for some $\tau^* \in (0, 0.5)$, let us assume that Assumptions (A1), (A2), (A3*), (A4), (A5) and (P) hold for some fixed $\alpha \geq 0$. Then, there exist MNPDPDEs $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_S^t, \widehat{\boldsymbol{\beta}}_N^t)^t$ of $\boldsymbol{\beta}$, with $\widehat{\boldsymbol{\beta}}_S \in \mathbb{R}^s$, such that $\widehat{\boldsymbol{\beta}}$ is a (strict) local minimizer (MPDPDE) and satisfies the following results with probability tending to 1 as $n \to \infty$.*

a) $\widehat{\boldsymbol{\beta}}_N = \mathbf{0}_{p-s}$, *and* $\left\| \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \right\|_2 = O(\sqrt{s/n})$.

b) *Let* $\boldsymbol{A}_n \in R^{q \times (s+1)}$ *such that* $\boldsymbol{A}_n \boldsymbol{A}_n^t \to \boldsymbol{G}$ *as* $n \to \infty$, *where* $\boldsymbol{G}$ *is symmetric and positive definite. Then,*

$$\boldsymbol{A_n \Omega_{n,\alpha}(\beta_0)^{-1/2} \Psi_{n,\alpha}(\beta_0)} \left(\widehat{\boldsymbol{\beta}}_{\mathbf{S}} - \boldsymbol{\beta}_{\mathbf{0S}} \right) \xrightarrow{\mathcal{D}} \mathbf{N_q}(\mathbf{0_q}, \boldsymbol{G}).$$

Outline

Penalized log-DPD or γ-divergence

The log-DPD or γ-divergence (Jones et al., 2001; Fujisawa and Eguchi, 2008)

$$d_\gamma(g, f_\theta) = \log \int f_\theta^{1+\gamma} - \frac{1+\gamma}{\gamma} \log \int f_\theta^\gamma g + \frac{1}{\gamma} \log \int g^{1+\gamma}, \quad \text{if} \quad \gamma > 0,$$

$$d_0(g, f_\theta) = \lim_{\gamma \to 0} d_\gamma(g, f_\theta) = \int g \log(g/f_\theta) = \text{KL-divergence}.$$

The log-DPD or γ-divergence loss function for LRM (Fujisawa and Eguchi, 2008)

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

$$\mathbf{L_n^{(\gamma)}}(\theta) = -\frac{\mathbf{1}}{\gamma} \log \left(\frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{f_\theta(y_i|x_i)} \right) + \frac{\mathbf{1}}{\mathbf{1+\gamma}} \log \left(\frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \int \mathbf{f_\theta(y|x_i)dy} \right), \quad \mathbf{f_\theta(\cdot|x_i)} \equiv \mathcal{N}(\mathbf{x_i^T}\boldsymbol{\beta}, \boldsymbol{\sigma^2}).$$

Minimum Penalized log-DPD Estimator $\left(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}\right) = \arg\min_{(\boldsymbol{\beta}, \sigma)} \left\{ L_n^{(\gamma)}(\boldsymbol{\beta}, \sigma) + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \right\}$

- Kawashima and Fujisawa (2017): Proposed and studied empirically with ℓ_1-penalty only! (no theory)
- Castilla et al. (2020): Derived general theory and IF for general non-concave penalties (similarly as DPD)!

Penalized Bregman divergence

The Bregman divergence (Bregman, 1967))

$$Q(\nu, \mu) = -q(\nu) + q(\mu) + (\nu - \mu)q'(\mu), \qquad q \text{ concave.} \tag{12}$$

The log-DPD or γ-divergence loss function for GLM (Fujisawa and Eguchi, 2008)

Given a covariate value $\mathbf{X} = \boldsymbol{x}$, the response variable Y has density

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

$$\mathbf{L_n^{(q)}}(\beta) = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{Q}(\mathbf{y_i}, \mathbf{g^{-1}}(\boldsymbol{x}_\mathbf{i}^\mathbf{T}\beta)).$$

Minimum Penalized Bregman Divergence Estimator (MPBDE)

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ L_n^{(q)}(\boldsymbol{\beta}) + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \right\}$$

- Zhang et al. (2010): Studied the properties of MPBDE for the cases $p < n$ and pn!
- Zhang et al. (2011): Developed an appropriate algorithm for computation of MPBDE with ℓ_1 penalty!
- Jiang and Zhang (2013): Studied the properties of MPBDE for the high-dimensional cases with $p > n$!

Outline

Simulation Setup and Performance Metrics

- 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 $\boldsymbol{\epsilon} \sim N(0, 0.5^2)$, 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.
- Methods compared- RLARS, sLTS, LAD-Lasso, DPD-lasso, log DPD-lasso, DPD-ncv(with SCAD), along with non-robust Lasso, SCAD and MCP.
- RLARS solution is used as our starting point.

Perfromance Metrics

$$\text{MSEE}(\widehat{\boldsymbol{\beta}}) = (1/p)\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2, \quad \text{RMSPE}(\widehat{\boldsymbol{\beta}}) = \sqrt{\|\mathbf{y}_{test} - \mathbf{X}_{test}\widehat{\boldsymbol{\beta}}\|^2},$$

$$\text{EE}(\widehat{\sigma}) = |\widehat{\sigma} - \sigma_0|, \quad \text{MS}(\widehat{\boldsymbol{\beta}}) = |\,\text{supp}(\widehat{\boldsymbol{\beta}})|.$$

$$\text{TP}(\widehat{\boldsymbol{\beta}}) = \frac{|\,\text{supp}(\widehat{\boldsymbol{\beta}}) \cap \text{supp}(\boldsymbol{\beta}_0)|}{|\,\text{supp}(\boldsymbol{\beta}_0)|}, \quad \text{TN}(\widehat{\boldsymbol{\beta}}) = \frac{|\,\text{supp}(\widehat{\boldsymbol{\beta}}) \cap \text{supp}(\boldsymbol{\beta}_0)|}{|\,\text{supp}(\boldsymbol{\beta}_0)|},$$

Table of outputs for $p = 500$, $n = 100$ and Y-outliers

Setting B: $\beta_1 = \beta_7 = 1.5$, $\beta_2 = 0.5$, $\beta_4 = \beta_{11} = 1$, and 0 otherwise.						
Method	MSEE($\widehat{\boldsymbol{\beta}}$) ($\times 10^{-4}$)	RMSPE($\widehat{\boldsymbol{\beta}}$) ($\times 10^{-2}$)	EE($\widehat{\sigma}$)	TP($\widehat{\boldsymbol{\beta}}$)	TN($\widehat{\boldsymbol{\beta}}$)	MS($\widehat{\boldsymbol{\beta}}$)
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
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

Setting B: $\beta_1 = \beta_7 = 1.5$, $\beta_2 = 0.5$, $\beta_4 = \beta_{11} = 1$, and 0 otherwise.						
Method	MSEE($\widehat{\boldsymbol{\beta}}$) ($\times 10^{-4}$)	RMSPE($\widehat{\boldsymbol{\beta}}$) ($\times 10^{-2}$)	EE($\widehat{\sigma}$)	TP($\widehat{\boldsymbol{\beta}}$)	TN($\widehat{\boldsymbol{\beta}}$)	MS($\widehat{\boldsymbol{\beta}}$)
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
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 $p = 500$, $n = 100$ and no outliers

Setting B: $\beta_1 = \beta_7 = 1.5$, $\beta_2 = 0.5$, $\beta_4 = \beta_{11} = 1$, and 0 otherwise.						
Method	MSEE($\widehat{\boldsymbol{\beta}}$) ($\times 10^{-4}$)	RMSPE($\widehat{\boldsymbol{\beta}}$) ($\times 10^{-2}$)	EE($\widehat{\sigma}$)	TP($\widehat{\boldsymbol{\beta}}$)	TN($\widehat{\boldsymbol{\beta}}$)	MS($\widehat{\boldsymbol{\beta}}$)
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
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

Outline

Summary

- We discussed the need minimum distance approach for robust estimation and variable selection in high-dimensional data analyses
- We discussed different minimum distance methods, namely the procedures based on L_2-divergence, DPD, log-DPD and Bregman divergence
- The DPD based procedures are discussed in great detail, including their IF, oracle consistency and asymptotic normality results, under high-dimensional linear regression model.
- Extension of the minimum penalized DPD estimator is also discussed for ultra-high dimensional GLM.
- Numerical illustrations are provided for comparison of different minimum distance methods, along with an interesting real data example.

References

- Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, **85**, 549–559.
- Basu, A., Shioya, H. and Park, C. (2011). *Statistical Inference: The Minimum Distance Approach*. Chapman & Hall/CRC, Boca de Raton.
- Castilla, E., Ghosh, A., Jaenada, M., and Pardo, L. (2020). On regularization methods based on Rényi's pseudodistances for sparse high-dimensional linear regression models. *arXiv preprint*, arXiv:2007.15929.
- Durio, A. and Isaia, E. D. (2011). The minimum density power divergence approach in building robust regression models. *Informatica*, **22**(1), 43–56.
- Fujisawa, H. and Eguchi, S. (2008). Robust Parameter Estimation with a Small Bias Against Heavy Contamination. *J. Mult. Anal.*, **99**, 2053–2081.
- Ghosh, A. (2021+). Robustness Concerns in High-dimensional Data Analysis and Potential Solutions. To appear in *Big Data Analytics in Chemoinformatics and Bioinformatics (with applications to computer-aided drug design, cancer biology, emerging pathogens and computational toxicology)*, Basak, S.C. and Vracko, M. Eds., Elsevier.
- 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. 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.

References

- Jiang, Y., and Zhang, C. (2013). High-dimensional regression and classification under a class of convex loss functions. *Stat. Its Interface*, **6(2)**, 285-299.
- Jones, M. C., Hjort, N. L., Harris, I. R., and Basu, A. (2001). A comparison of related density-based minimum divergence estimators. *Biometrika*, **88(3)**, 865–873.
- Kawashima, T. and Fujisawa, H. (2017). Robust and Sparse Regression via γ-Divergence. *Entropy*, **19(11)**, 608.
- Lozano, A. C. ; Meinshausen, N- and Yang, E. (2016). Minimum Distance Lasso for robust high-dimensional regression. *Electron. J. Stat.*, **10**, 1296–1340.
- Pardo, L. (2006). *Statistical inference based on divergence measures*. CRC press.
- Yuille, A. L., and Rangarajan, A. (2003). The concave-convex procedure. *Neural Comput.*, **15(4)**, 915–936.
- Zang, Y., Zhao, Q., Zhang, Q., et al. (2017). Inferring gene regulatory relationships with a high-dimensional robust approach. *Genet. Epidemiol.*, **41(5)**, 437–454.
- Zhang, C., Jiang, Y., and Chai, Y. (2010). Penalized Bregman divergence for large-dimensional regression and classification. *Biometrika*, **97(3)**, 551–566.
- Zhang, C., Zhang, Z., and Chai, Y. (2011). Penalized Bregman divergence estimation via coordinate descent. *JIRSS*, **10(2)**, 125–140

THANK YOU!

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