Robust group and simultaneous inferences for high-dimensional single index model

Weichao Yang*[1], Hongwei Shi*[1], Xu Guo✉[1], Changliang Zou[2]

[1]Beijing Normal University [2]Nankai University *contribute equally ✉corresponding author

Article link: https://neurips.cc/virtual/2024/poster/95500

Introduction

Consider the following general high-dimensional single index model (SIM):

$$Y = g(\beta^\top X, \epsilon) \quad \text{with} \quad \epsilon \perp X.$$

- The link function $g(\cdot)$ is unknown.
- $\beta = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$, and p **can be much larger than sample size** n.
- This model covers linear models, generalized linear models and classical SIM.

Introduction

We are interested in the following problems:

- **Group inference problem**

$$\mathbb{H}_{0,\mathcal{G}} : \beta_j = 0 \text{ for all } j \in \mathcal{G} \quad \text{versus} \quad \mathbb{H}_{1,\mathcal{G}} : \beta_j \neq 0 \text{ for some } j \in \mathcal{G}.$$

  Where $\mathcal{G}$ is a prespecified subset of $\{1, 2, \ldots, p\}$ with $p_0 = |\mathcal{G}|$.

- **Simultaneous inference problem**

$$\mathbb{H}_{0j} : \beta_j = 0 \quad \text{versus} \quad \mathbb{H}_{1j} : \beta_j \neq 0 \quad \text{for } 1 \leq j \leq p.$$

Applications: for example, in genome-wide association studies:

- Variant sets analysis.
- Identifying specific genes.

Our contributions

- A high-dimensional robust inference framework.
- Extension of the rank-LASSO procedure in Rejchel and Bogdan (2020).
- Asymptotically honest group inference procedure.
- Multiple testing procedure controlling false discovery rate (FDR).

Robust inference framework

Consider the pseudo-linear model:

$$h(Y) = \beta_h^\top X + e.$$

- $h(\cdot)$ is a transformation function.
- $\beta_h = (\beta_{h1}, \ldots, \beta_{hp})^\top$.
- The error term e satisfies $E(eX) = 0$.
- Under linear condition, β_h **is proportional to** β**.**
- We can recast the general SIM into a pseudo-linear model.

Robust inference framework

Problems aforementioned can be transformed as:

- **Group inference problem**

$$\mathbb{H}'_{0,\mathcal{G}} : \beta_{hj} = 0 \text{ for all } j \in \mathcal{G} \quad \text{versus} \quad \mathbb{H}'_{1,\mathcal{G}} : \beta_{hj} \neq 0 \text{ for some } j \in \mathcal{G}.$$

- **Simultaneous inference problem**

$$\mathbb{H}'_{0j} : \beta_{hj} = 0 \quad \text{versus} \quad \mathbb{H}'_{1j} : \beta_{hj} \neq 0 \quad \text{for } 1 \leq j \leq p.$$

For robust consideration, we consider the **distribution transformation of Y**, denote as $F(Y)$.

Group inference procedure

For each individual $\mathbb{H}'_{0j} : \beta_{hj} = 0$, define the standardized test statistic

$$\widetilde{T}_{nj} = \frac{1}{\hat{\sigma}_j \sqrt{n}} \sum_{i=1}^{n} \left\{ F_n(Y_i) - 1/2 - Z_{ij}^\top \hat{\gamma}_j \right\} (X_{ij} - Z_{ij}^\top \hat{\theta}_j),$$

- $F_n(Y_i) = n^{-1} \sum_{j=1}^{n} I(Y_j \leq Y_i)$ and $\hat{\sigma}_j^2$ is the estimator of variance.
- $\hat{\gamma}_j$ is subvector of $\hat{\beta}_h$ without $\hat{\beta}_{hj}$, $p_{\lambda_Y}(\cdot)$ is the penalty function and

$$\hat{\beta}_h = \arg \min_{\beta_h \in \mathbb{R}^p} (2n)^{-1} \sum_{i=1}^{n} \left\{ F_n(Y_i) - 1/2 - X_i^\top \beta_h \right\}^2 + \sum_{l=1}^{p} p_{\lambda_Y}(|\beta_{hl}|).$$

- Z_{ij} is subvector of X_i without X_{ij}, $p_{\lambda_X}(\cdot)$ is the penalty function and

$$\hat{\theta}_j = \arg \min_{\theta_j \in \mathbb{R}^{p-1}} \frac{1}{2n} \sum_{i=1}^{n} \left(X_{ij} - Z_{ij}^\top \theta_j \right)^2 + \sum_{l=1}^{p-1} p_{\lambda_X}(|\theta_{jl}|),$$

Group inference procedure

For group inference problem, we consider the maximum type test statistic. That is,

$$M_{n,\mathcal{G}} = \max_{j \in \mathcal{G}} \widetilde{T}_{nj}^2.$$

We can reject the null hypothesis $\mathbb{H}_{0,\mathcal{G}}$ at the significant level α if and only if

$$M_{n,\mathcal{G}} \geq c_{\mathcal{G}}(\alpha),$$

where $c_{\mathcal{G}}(\alpha) = 2\log p_0 - \log\log p_0 + q_\alpha$ and

$$q_\alpha = -\log(\pi) - 2\log\log(1-\alpha)^{-1}.$$

Simultaneous inference procedure

Define $\mathcal{H}_0 = \{j : \beta_j = 0,\ j = 1, \ldots, p\}$. At a given threshold level $t > 0$, $\mathbb{H}_{0j}$ is rejected if $|\widetilde{T}_{nj}| \geq t$. Accordingly, the false discovery proportion (FDP) and FDR are

$$\mathrm{FDP}(t) = \frac{\sum_{j \in \mathcal{H}_0} I\big(|\widetilde{T}_{nj}| \geq t\big)}{\max\big\{\sum_{j=1}^{p} I\big(|\widetilde{T}_{nj}| \geq t\big), 1\big\}}, \qquad \mathrm{FDR}(t) = \mathbb{E}\{\mathrm{FDP}(t)\}.$$

As $\mathcal{H}_0$ is unknown, we **use $pG(t)$ to approximate** $\sum_{j \in \mathcal{H}_0} I(|\widetilde{T}_{nj}| \geq t)$, where $G(t) = 2 - 2\Phi(t)$.

Simultaneous inference procedure

In summary, we have the following procedure controlling the FDR and FDP at a pre-specified level $0 < \alpha < 1$:

1. Let $b_p = \sqrt{2 \log p - \log \log p}$ and define

$$\hat{t} = \inf\left\{ 0 \le t \le b_p : \frac{pG(t)}{\max\{\sum_{j=1}^{p} I(|\widetilde{T}_{nj}| \ge t), 1\}} \le \alpha \right\}.$$

2. If $\hat{t}$ does not exist, then let $\hat{t} = \sqrt{2 \log p}$. We reject $\mathbb{H}_{0j}$ whenever $|\widetilde{T}_j| \ge \hat{t}$.

Application to Bacillus Subtilis dataset

Description of the dataset

- $n = 71$ observations of strains of Bacillus Subtilis.
- $p = 300$ covariates, measuring the log-expression levels of 300 selected genes.
- The response variable is logarithm of the riboflavin production rate.

We are interested in

- Detect which genes are associated with riboflavin production rate.
  - 10 significant genes with the FDR level of 0.1, denote as $\mathcal{G}_1$.
  - 15 significant genes with the FDR level of 0.2, denote as $\mathcal{G}_2$.
- Test whether there exist other significant genes in $\mathcal{G}_1^c$ or $\mathcal{G}_2^c$.
  - P-values correspond to $\mathcal{G}_1^c$ and $\mathcal{G}_2^c$ are 0.727 and 0.937.
  - No significant gene in $\mathcal{G}_1^c$ and $\mathcal{G}_2^c$.

Thanks for listening

NEURAL INFORMATION
PROCESSING SYSTEMS

References I

Rejchel, W. and M. Bogdan (2020). Rank-based lasso - efficient methods for high-dimensional robust model selection. *Journal of Machine Learning Research 21*(244), 1–47.