A General Framework for Sparse Sufficient Dimension Reduction

Jing Zeng

Department of Statistics
Florida State University

Joint work with
Dr. Xin Zhang and **Dr. Qing Mai**
July 29th, 2019

Motivation

Motivation

- High-dimensional dataset: sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}$ is singular.
- Dimension selection, variable selection and central subspace estimation
  - Existing methods: **multi-stage**
  - Our method: **simultaneously**
- Our method estimates a $p \times H$ matrix and $d < H \ll p$, where d is the dimension of central subspace.
- General framework: applied to existing sufficient dimension reduction methods.

Introduction

Sufficient dimension reduction (SDR) and central subspace (CS)

Variables:

- $Y \in \mathbb{R}$: **response**
- $\mathbf{X} \in \mathbb{R}^p$: **predictor**

Definition:

- A subspace $\mathrm{span}(\mathbf{B})$, the span of the basis matrix $\mathbf{B} \in \mathbb{R}^{p \times d}$, $d < p$, is called a **central subspace (CS)** if it is the smallest subspace such that

$$Y \perp\!\!\!\perp \mathbf{X} | \mathbf{B}^T \mathbf{X} \quad \Longleftrightarrow \quad Y|\mathbf{X} \sim Y|\mathbf{B}^T\mathbf{X}.$$

  And it is denoted by $\mathcal{S}_{Y|\mathbf{X}}$.
- By projecting $\mathbf{X}$ onto $\mathrm{span}(\mathbf{B})$, the dimension is reduced from p to d, which is called **Sufficent Dimension Reduction (SDR)**.

Sliced Inverse Regression (SIR)

- Partition the range of Y into disjoint intervals, each interval is called the **slice** of Y.
- Assume

$$\mathcal{S}_{Y|\mathbf{X}} = \mathbf{\Sigma}_{\mathbf{X}}^{-1}\mathsf{span}\{\mu_1 - \overline{\mu}, \dots, \mu_H - \overline{\mu}\} = \mathsf{span}(\mathbf{\Sigma}_{\mathbf{X}}^{-1}\mathbf{U}),$$

  - $\mathbf{U} = (\mu_1 - \overline{\mu}, \dots, \mu_H - \overline{\mu}) \in \mathbb{R}^{p \times H}$
  - $\mu_h = \mathbb{E}[\mathbf{X}|J_h(Y) = 1]$: the h-th within-slice mean

$$J_h(y) = \begin{cases} 1, & y \text{ is in the } h\text{-th slice}, \\ 0, & \text{otherwise}, \end{cases} \qquad h = 1, \dots, H$$

  - $\overline{\mu} = \mathbb{E}[\mathbf{X}]$: the marginal mean of $\mathbf{X}$
- Estimate $\mathcal{S}_{Y|\mathbf{X}}$ by estimating $p \times H$ matrix $\mathbf{\Sigma}_{\mathbf{X}}^{-1}\mathbf{U}(H \ll p)$.
- $\text{rank}(\mathcal{S}_{Y|\mathbf{X}}) = \text{rank}(\mathbf{\Sigma}_{\mathbf{X}}^{-1}\mathbf{U}) = d < H$

Methods

Sparse Sufficient Dimension Reduction (SSDR)

- **Objective function:**

$$L(\mathbf{B}) = \frac{1}{2}\mathrm{tr}\left(\mathbf{B}^{\top}\mathbf{M}\mathbf{B}\right) - \mathrm{tr}\left(\mathbf{U}^{\top}\mathbf{B}\right).$$

where $\mathbf{M} \in \mathbb{R}^{p\times p} > 0$, $\mathbf{U} \in \mathbb{R}^{p\times H}(H \ll p)$.

$$\mathbf{B}^{*} = \operatorname*{argmin}_{\mathbf{B}\in\mathbb{R}^{p\times H}} L(\mathbf{B}) = \mathbf{M}^{-1}\mathbf{U}$$

Penalty terms

- **Penalized objective function**:

$$\widehat{S}(\mathbf{B};\lambda_1,\lambda_2) = \frac{1}{2}\mathrm{tr}\left(\mathbf{B}^T\widehat{\mathbf{M}}\mathbf{B}\right) - \mathrm{tr}\left(\widehat{\mathbf{U}}^T\mathbf{B}\right) + \lambda_1\|\mathbf{B}\|_{2,1} + \lambda_2\|\mathbf{B}\|_*$$

- $\|\mathbf{B}\|_{2,1} = \sum_{i=1}^{p}\sqrt{\sum_{j=1}^{H} b_{i,j}^2}$ (impose **sparsity**).
- $\|\mathbf{B}\|_* = \sum_{i=1}^{H}\sigma_i$, where σ_i is the i-th singular value (impose **low-rank structure**).
- **Goal:** optimize the penalized objective function:

$$\widehat{\mathbf{B}} = \underset{\mathbf{B}\in\mathbb{R}^{p\times H}}{\mathrm{argmin}}\ \hat{S}(\mathbf{B};\lambda_1,\lambda_2).$$

$\widehat{\mathbf{B}}$ is the sparse and low-rank estimation of $\mathbf{M}^{-1}\mathbf{U}$.

Examples

- Sliced inverse regression (SIR) [Li, 1991]

$$\widehat{\mathbf{M}} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}, \widehat{\mathbf{U}} = (\hat{\mu}_1 - \hat{\mu}, \cdots, \hat{\mu}_H - \hat{\mu})$$

- Intraslice covariance [Cook and Ni, 2005]

$$\widehat{\mathbf{M}} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}, \widehat{\mathbf{U}} = \left(\widehat{\mathrm{Cov}}(\mathbf{X}, YJ_1(Y)), \ldots, \widehat{\mathrm{Cov}}(\mathbf{X}, YJ_H(Y))\right)$$

$$J_h(y) = \begin{cases} 1, & y \text{ is in the } h\text{-th slice}, \\ 0, & \text{otherwise.} \end{cases} \quad h = 1, \ldots, H$$

- Principle Fitted Components (PFC) [Cook and Forzani, 2008]

$$\widehat{\mathbf{M}} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}, \widehat{\mathbf{U}} = \tilde{\mathbf{X}}^T \tilde{\mathbf{F}} (\tilde{\mathbf{F}}^T \tilde{\mathbf{F}})^{-1/2}$$

  — The ith row of $\tilde{\mathbf{X}} = \mathbf{X}_i - \overline{\mathbf{X}}$, where $\mathbf{X}_i$ is the i-th sample.
  — The ith row of $\tilde{\mathbf{F}} = \mathbf{f}(y_i) - \bar{\mathbf{f}}$, where $\mathbf{f}(y_i)$ is a vector of functions of sample y_i.

Algorithm: ADMM

- Recast the problem

$$\widehat{\mathbf{B}} = \underset{\mathbf{B}\in\mathbb{R}^{p\times H}}{\operatorname{argmin}} \widehat{S}(\mathbf{B};\lambda_1,\lambda_2)$$

to

$$\begin{aligned} &\underset{\mathbf{B}\in\mathbb{R}^{p\times H}}{\operatorname{argmin}} && \frac{1}{2}\operatorname{tr}\left(\mathbf{B}^{\mathsf{T}}\widehat{\mathbf{M}}\mathbf{B}\right) - \operatorname{tr}\left(\widehat{\mathbf{U}}^{\mathsf{T}}\mathbf{B}\right) + \lambda_1\|\mathbf{B}\|_{2,1} + \lambda_2\|\mathbf{C}\|_*, \\ &\text{subject to} && \mathbf{B} = \mathbf{C}. \end{aligned}$$

- Implement the **ADMM** algorithm.

Parameter tuning

Method: cross-validation

- For each combination of $(\lambda_1, \lambda_2, \gamma)$,

$$(\mathsf{Y}_{\mathsf{train}}, \mathbf{X}_{\mathsf{train}}) \xrightarrow{\text{SSDR-based method}} \widehat{\mathbf{B}}$$

- Projecting $\mathbf{X}_{\mathsf{val}}$ onto $\widehat{\mathbf{B}}$:

$$\mathbf{X}_{\mathsf{val}} \xrightarrow{\text{projection}} \tilde{\mathbf{X}} := \widehat{\mathbf{B}}^{\mathsf{T}} \mathbf{X}_{\mathsf{val}}$$

- Fit (linear) regression model:

$$(\mathsf{Y}_{\mathsf{val}}, \tilde{\mathbf{X}}) \xrightarrow{\text{(linear) regression}} \text{RMSE}$$

- Minimal RMSE $\Leftrightarrow$ optimal $(\lambda_1, \lambda_2, \gamma)$.

Numerical Study

Comparison between SSDR-based methods and other competitor

Our methods

- SSDR-SIR
- SSDR-intra
- SSDR-PFC

Competitor:

- LassoSIR[1]

[1] Qian Lin, Zhigen Zhao and Jun S. Liu. Sparse Sliced Inverse Regression via Lasso. *Journal of the American Statistical Association*, 0(0):1-33, 2019.

Simulation models

- $X \in \mathbb{R}^p \sim N(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = AR(0.5)$.
- $\epsilon \sim N(0, 1)$
- $p = 1000, N = 500$
- d : dimension of central subspace
- s : the number of non-zero variables in rows
- **Models**:
  - Model I (linear model, d = 1, s =20):

$$y = \mathbf{B}^T X + 0.5\epsilon,\ \ \mathbf{B} \in \mathbb{R}^p.$$

  - Model II (single index model, d =1, s = 20):

$$y = (\mathbf{B}^T X)^3/2 + \epsilon,\ \ \mathbf{B} \in \mathbb{R}^p.$$

  - Model III(1~3) (multiple index model, d = 2, s = 6):

$$y = (\beta_1^T X) \cdot \exp(\beta_2^T X) + \sigma \cdot \epsilon,\ \ \mathbf{B} = (\beta_1, \beta_2) \in \mathbb{R}^{p \times 2}$$

    where $\sigma = 0.2, 0.6, 1$.

Subspace distance

- **Subspace distance:** For the true central subspace $\mathbf{B}$ and the estimator $\widehat{\mathbf{B}}$,

$$D(\mathbf{B}, \widehat{\mathbf{B}}) = \|\mathbf{P_B} - \mathbf{P_B}\mathbf{P}_{\widehat{\mathbf{B}}}\mathbf{P_B}\|_F/\sqrt{d},$$

where d is the dimension of true subspace $\mathbf{B}$.

Model	SSDR-SIR	SSDR-intra	SSDR-PFC	LassoSIR
I	0.1(0.29)	**0.08(0.29)**	0.09(0.3)	0.12(0.35)
II	0.17(0.72)	**0.16(0.67)**	0.16(0.64)	0.39(0.88)
III(1)	0.19(1.89)	0.21(1.91)	**0.14(1.41)**	0.36(2.05)
III(2)	0.25(2.2)	**0.22(1.92)**	0.28(2.58)	0.57(1.46)
III(3)	0.25(1.99)	**0.24(1.99)**	0.33(2.49)	0.68(0.75)

Table: Mean subspace distance $D(\mathbf{B}, \widehat{\mathbf{B}})$ and standard error $(\times 10^{-2})$

Rank estimation

Model	SSDR-SIR	SSDR-intra	SSDR-PFC	LassoSIR
I(d=1)	**100**	95	92	76
II(d=1)	76	**89**	67	76
III(1)(d=2)	76	**89**	66	88
III(2)(d=2)	72	**90**	54	69
III(3)(d=2)	63	**86**	52	47

Table: The correct rate (%) of rank estimation

Variable selection

Model	SSDR-SIR	SSDR-intra	SSDR-PFC	LassoSIR
I	99.5/0.03	99.65/0.03	99.15/**0.02**	99.95/5.65
II	95.95/0.71	97.4/**0.28**	95.6/0.74	93.55/7.96
III(1)	100/**0.27**	98.5/0.42	100/0.52	100/11.79
III(2)	100/**0.28**	99.83/0.47	100/0.3	100/10.17
III(3)	99.8/0.18	99/0.4	99.8/**0.13**	99.3/7.01

Table: True positive rate (%)/false positive rate (%). The standard errors are all less than 0.01.

References

R Dennis Cook and Liliana Forzani. Principal fitted components for dimension reduction in regression. *Statistical Science*, 23(4):485–501, 2008.

R Dennis Cook and Liqiang Ni. Using intraslice covariances for improved estimation of the central subspace in regression. *Biometrika*, 93(1):65–74, 2005.

Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

Thank you!