Functional and Very High Dimension Reduction

Yanyuan Ma

Penn State University

joint work with Fei Jiang, Seungchul Baek, Jiguo Cao, Ying Wei

Outline

- Traditional dimension reduction and literature
- Two related extensions
  - Functional dimension reduction
  - Very high dimension reduction
- Motivation
- Model and estimation procedure
- Theoretic properties and numeric evaluations
- Data analysis

Traditional Dimension Reduction

1. Distribution of Y relates to covariates x only through $\beta^{\mathrm{T}} x$.

$$Y \perp\!\!\!\perp x \mid \beta^{\mathrm{T}} x$$
$$\text{or} \quad \mathrm{pr}(Y \leq y \mid x) = \mathrm{pr}(Y \leq y \mid \beta^{\mathrm{T}} x) \text{ for all } y.$$

   Central space $\mathcal{S}_{Y|X}$: Span of the columns in β.
2. Mean of Y relates to covariates x only through $\beta^{\mathrm{T}} x$.

$$E(Y \mid x) = E(Y \mid \beta^{\mathrm{T}} x).$$

   Central mean space $\mathcal{S}_{E(Y|X)}$: Span of the columns in β.
3. Specifics: $Y \in R$, $x \in R^{p \times 1}$, $\beta \in R^{p \times d}$, $p > d$.
4. Goal: Estimate $\mathcal{S}_{Y|X}$ or $\mathcal{S}_{E(Y|X)}$.

Three Classes of Estimation Approaches

1. Inverse regression based methods.
   - $\mathcal{S}_{Y|X}$: SIR (Li 1991), SAVE (Cook and Weisberg 1991), DR (Li and Wang 2007)
   - $\mathcal{S}_{E(Y|X)}$: OLS (Li and Duan 1989), PHD (Li 1992, Cook and Li 2002)
2. Nonparametric estimation:
   dMAVE (Xia 2007), MAVE (Xia, Tong, Li and Zhu 2002)
3. Semiparametric estimation:
   - $\mathcal{S}_{Y|X}$: Semi-SIR, Semi-DR, Semi-SAVE (Ma and Zhu 2012), Efficient (Ma and Zhu 2013)
   - $\mathcal{S}_{E(Y|X)}$: Semi-PHD (Ma and Zhu 2012), Local Efficient (Ma and Zhu 2014)

Functional Dimension Reduction

Pollution and Cardiovascular Disease

- Is air pollution a risk factor for cardiovascular disease (CVD) occurrence and death?
- NMMAPSdata (EPA/United States Census Bureau)
- Pollutants: carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), ozone (O_3). Coded as $X_1(t), \ldots, X_4(t)$.
- Daily record of four pollutants and CVD death rate, 108 US cities, 1987-2000, many missing. Annual CVD death rate (Y).
- Cleaned out 400 observations, each over a course of one year.
- Goal: study how the pollution ($X(t)$) of a city in one year affect the city's death rate (Y) in the following year.

Model

$$f_{Y|X(t)}\{Y, X(t)\} = f\left\{Y, \int_0^1 \beta^{\mathrm{T}} X(t)\alpha(t)dt\right\}$$

- $Y \in \mathcal{R}$, $X(t) \in \mathcal{R}^{p\times 1}$, $\beta \in \mathcal{R}^p$, $\alpha(t) \in \mathcal{R}$.
- Pollutant effect combined through β.
- Seasonal effect regulated through $\alpha(t)$.
- Dependence of CVD death on the pollutant effect summary f unspecified.
- If $\alpha(t)$ known, then sufficient dimension reduction model: $f_{Y|X(t)}\{Y, X(t)\} = f\left\{Y, \beta^{\mathrm{T}} \int_0^1 X(t)\alpha(t)dt\right\}$.
- If β known, then functional single index model: $f_{Y|X(t)}\{Y, X(t)\} = f\left\{Y, \int_0^1 \beta^{\mathrm{T}} X(t)\alpha(t)dt\right\}$.

Model after B-spline Approximation of $\alpha(t)$

- $\alpha(t) \approx \mathbf{B}_r(t)^{\mathrm{T}}\gamma$, $\mathbf{B}_r(t) \in \mathcal{R}^q, \gamma \in \mathcal{R}^q$.
- Model is approximated by

$$f\left\{Y, \beta^{\mathrm{T}} \int_0^1 X(t)\alpha(t)dt\right\} \approx f\left\{Y, \beta^{\mathrm{T}} \int_0^1 X(t)\mathbf{B}_r(t)^{\mathrm{T}}dt\gamma\right\}.$$

- Let $\mathbf{Z} = \int_0^1 X(t)\mathbf{B}_r(t)^{\mathrm{T}}dt \in \mathcal{R}^{p\times q}$.
- New model (dimension folding)

$$f_{Y|X(t)}\{Y, X(t)\} \approx f(Y, \beta^{\mathrm{T}}\mathbf{Z}\gamma)$$

- Estimating $\beta, \alpha(t)$ via estimating β, γ.
- When $q \to \infty$, $\mathbf{B}_r^{\mathrm{T}}(t)\gamma$ approximates $\alpha(t)$ sufficiently well.

Estimation of Dimension Folding Model

$$f_{Y|\mathbf{Z}}(Y, \mathbf{Z}, \beta, \gamma) = f(Y, \beta^{\mathrm{T}}\mathbf{Z}\gamma)$$

- View f as nuisance parameter, it is a semiparametric model.
- $\Lambda^{\perp} = \{\mathbf{a}(Y, \mathbf{Z}) - E(\mathbf{a} \mid Y, \beta^{\mathrm{T}}\mathbf{Z}\gamma) : E(\mathbf{a} \mid \mathbf{Z}) = E(\mathbf{a} \mid \beta^{\mathrm{T}}\mathbf{Z}\gamma)\}$
- $\{\mathbf{a}(\mathbf{Z}) - E(\mathbf{a} \mid \beta^{\mathrm{T}}\mathbf{Z}\gamma)\}\{\mathbf{b}(Y, \beta^{\mathrm{T}}\mathbf{Z}\gamma) - E(\mathbf{b} \mid \beta^{\mathrm{T}}\mathbf{Z}\gamma)\} \in \Lambda^{\perp}$
- $\sum_j \{\mathbf{a}_j(\mathbf{Z}) - E(\mathbf{a}_j \mid \beta^{\mathrm{T}}\mathbf{Z}\gamma)\}\{\mathbf{b}_j(Y, \beta^{\mathrm{T}}\mathbf{Z}\gamma) - E(\mathbf{b}_j \mid \beta^{\mathrm{T}}\mathbf{Z}\gamma)\} \in \Lambda^{\perp}$
- (Local) efficient score

$$\begin{bmatrix} \{\mathbf{Z} - E(\mathbf{Z} \mid \beta^{\mathrm{T}}\mathbf{Z}\gamma)\}\gamma \\ \{\mathbf{Z} - E(\mathbf{Z} \mid \beta^{\mathrm{T}}\mathbf{Z}\gamma)\}^{\mathrm{T}}\beta \end{bmatrix} \times \left[\frac{\partial \log f^*(Y, \beta^{\mathrm{T}}\mathbf{Z}\gamma)}{\partial(\beta^{\mathrm{T}}\mathbf{Z}\gamma)} - E\left\{ \frac{\partial \log f^*(Y, \beta^{\mathrm{T}}\mathbf{Z}\gamma)}{\partial(\beta^{\mathrm{T}}\mathbf{Z}\gamma)} \mid \beta^{\mathrm{T}}\mathbf{Z}\gamma \right\} \right]$$

Properties and Implementation

$$\{\mathbf{a}(\mathbf{Z}) - E(\mathbf{a} \mid \beta^{\mathrm{T}}\mathbf{Z}\gamma)\}\{\mathbf{b}(Y, \beta^{\mathrm{T}}\mathbf{Z}\gamma) - E(\mathbf{b} \mid \beta^{\mathrm{T}}\mathbf{Z}\gamma)\}$$

- Estimate $E(\mathbf{a} \mid \beta^{\mathrm{T}}\mathbf{Z}\gamma)$ and/or $E(\mathbf{b} \mid \beta^{\mathrm{T}}\mathbf{Z}\gamma)$ via kernel smoothing.

$$\begin{aligned}
\widehat{E}(\mathbf{a} \mid \beta^{\mathrm{T}}\mathbf{Z}\gamma) &= \frac{\sum_{i=1}^{n} \mathbf{a}(\mathbf{Z}_i) K_h(\beta^{\mathrm{T}}\mathbf{Z}_i\gamma - \beta^{\mathrm{T}}\mathbf{Z}\gamma)}{\sum_{i=1}^{n} K_h(\beta^{\mathrm{T}}\mathbf{Z}_i\gamma - \beta^{\mathrm{T}}\mathbf{Z}\gamma)}, \\
\widehat{E}(\mathbf{b} \mid \beta^{\mathrm{T}}\mathbf{Z}\gamma) &= \frac{\sum_{i=1}^{n} \mathbf{b}(Y_i, \beta^{\mathrm{T}}\mathbf{Z}_i\gamma) K_h(\beta^{\mathrm{T}}\mathbf{Z}_i\gamma - \beta^{\mathrm{T}}\mathbf{Z}\gamma)}{\sum_{i=1}^{n} K_h(\beta^{\mathrm{T}}\mathbf{Z}_i\gamma - \beta^{\mathrm{T}}\mathbf{Z}\gamma)}.
\end{aligned}$$

- Estimating both provides double insurance due to double robustness.
- Obtain $\widehat{\beta}, \widehat{\gamma}$ from

$$\sum_{i=1}^{n} \{\mathbf{a}(\mathbf{Z}_i) - \widehat{E}(\mathbf{a} \mid \beta^{\mathrm{T}}\mathbf{Z}_i\gamma)\}\{\mathbf{b}(Y_i, \beta^{\mathrm{T}}\mathbf{Z}_i\gamma) - \widehat{E}(\mathbf{b} \mid \beta^{\mathrm{T}}\mathbf{Z}_i\gamma)\} = 0$$

Efficient Estimator

Obtain $\widehat{\beta}_{\text{eff}}$, $\widehat{\gamma}$ from solving

$$\sum_{i=1}^{n} \begin{bmatrix} \{\mathbf{Z}_i - \widehat{E}(\mathbf{Z}_i \mid \beta^{\mathrm{T}}\mathbf{Z}_i\gamma)\}\gamma \\ \{\mathbf{Z}_i - \widehat{E}(\mathbf{Z}_i \mid \beta^{\mathrm{T}}\mathbf{Z}_i\gamma)\}^{\mathrm{T}}\beta \end{bmatrix}$$
$$\times \left[\frac{\partial \log \widehat{f}(Y_i, \beta^{\mathrm{T}}\mathbf{Z}_i\gamma)}{\partial(\beta^{\mathrm{T}}\mathbf{Z}_i\gamma)} - \widehat{E}\left\{ \frac{\partial \log \widehat{f}(Y_i, \beta^{\mathrm{T}}\mathbf{Z}_i\gamma)}{\partial(\beta^{\mathrm{T}}\mathbf{Z}_i\gamma)} \mid \beta^{\mathrm{T}}\mathbf{Z}_i\gamma \right\} \right] = 0,$$

where $\widehat{f}(Y, \beta^{\mathrm{T}}\mathbf{Z}_i\gamma) = c_0, \partial\widehat{f}(Y, \beta^{\mathrm{T}}\mathbf{Z}_i\gamma)/\partial(\beta^{\mathrm{T}}\mathbf{Z}_i\gamma) = c_1$ from minimizing

$$\sum_{j=1}^{n} \{K_b(Y_j - Y) - c_0 - c_1(\beta^{\mathrm{T}}\mathbf{Z}_j\gamma - \beta^{\mathrm{T}}\mathbf{Z}_i\gamma)\}^2 K_h(\beta^{\mathrm{T}}\mathbf{Z}_j\gamma - \beta^{\mathrm{T}}\mathbf{Z}_i\gamma).$$

Form $\widehat{\alpha}(t) = \mathbf{B}_r^{\mathrm{T}}(t)\widehat{\gamma}$.

Asymptotic Properties

Results established:

- $\widehat{\boldsymbol{\beta}}_{\text{eff}}$: Consistency, normality, efficiency, estimation variance.
- $\widehat{\alpha}(t)$: Bias, variance, normality.

Difficulties overcome:

- $\mathbf{Z}$ is a growing matrix.
- B-spline approximation imbedded in the kernel estimation.
- $\widehat{\boldsymbol{\beta}}$ is established under the original model, which has functional covariate and parameter.
- Efficient variance of $\widehat{\boldsymbol{\beta}}_{\text{eff}}$ is not the usual $\text{var}\{S_{\text{eff}}(\mathbf{Z}, Y, \boldsymbol{\beta}, \boldsymbol{\gamma})\}^{-1}$, additional projection needed in theoretical derivation and achieved via B-spline approximation in implementation.

Simulations

- Simulation 1

$$p = 9, n = 500, \alpha(t) = \sin(\pi t) + 1,$$
$$X \sim \text{uniform}(-5, 5),$$
$$Y \sim N(\int_0^1 \beta^{\mathrm{T}} x(t)\alpha(t)dt, 1)$$

- Simulation 2

$$p = 4, n = 500,$$
$$\alpha(t) = \text{polynomial of order 6},$$
$$X = \text{quadratic functions of } t + \text{ uniform noise},$$
$$Y \sim N(0.7269 + 0.53 \int_0^1 \beta^{\mathrm{T}} x(t)\alpha(t)dt, 0.05)$$

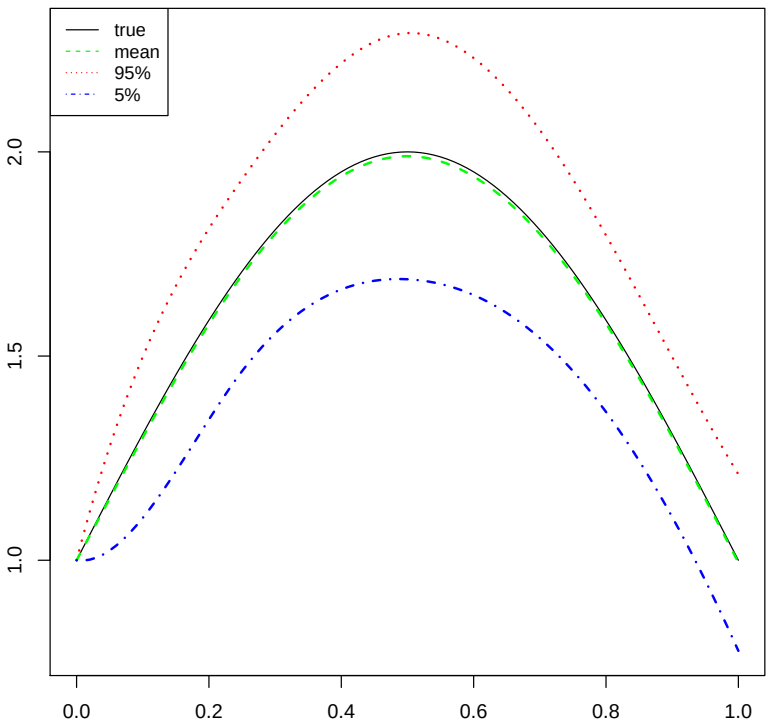
Simulation Results

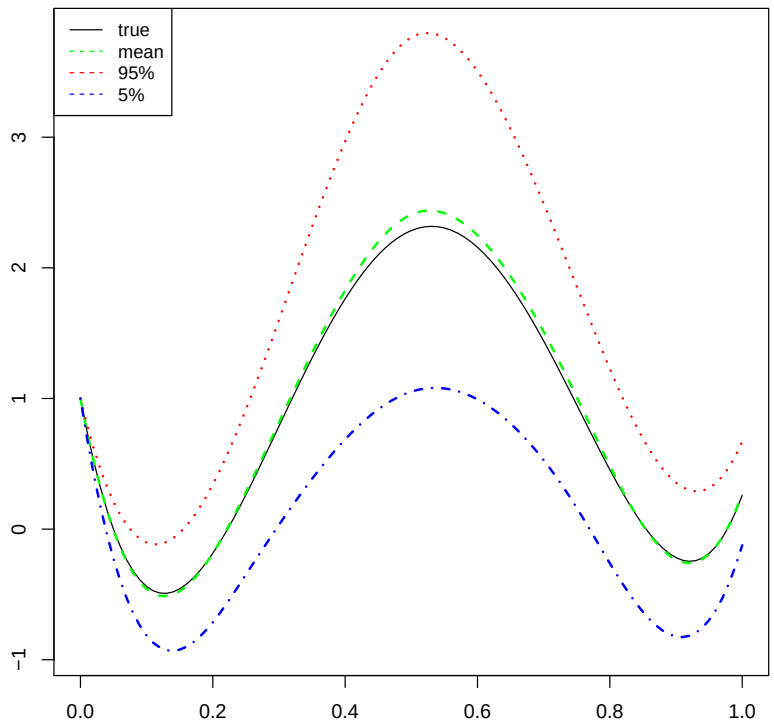
	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
	1	1.2	1.5	0.5	-0.5	-1.5	-1.2	-1
				Ora				
AE	1.01	1.21	1.51	0.51	-0.50	-1.51	-1.21	-1.01
SD	0.13	0.14	0.16	0.10	0.10	0.16	0.14	0.13
$\widehat{\text{SD}}$	0.12	0.14	0.16	0.10	0.10	0.16	0.14	0.12
CI	94.7	95.2	95.4	94.2	94.7	95.3	94.2	93.3
				Eff				
AE	1.03	1.23	1.54	0.52	-0.51	-1.53	-1.24	-1.02
SD	0.14	0.15	0.17	0.11	0.11	0.17	0.15	0.14
$\widehat{\text{SD}}$	0.14	0.15	0.17	0.11	0.11	0.17	0.15	0.14
CI	96.4	96.5	96.6	95.0	95.1	96.3	95.8	95.2
				Loc				
AE	1.03	1.24	1.55	0.52	-0.51	-1.55	-1.24	-1.03
SD	0.15	0.17	0.20	0.12	0.12	0.19	0.17	0.16
$\widehat{\text{SD}}$	0.15	0.17	0.20	0.12	0.12	0.20	0.17	0.15
CI	95.5	95.4	95.8	95.2	95.7	96.4	95.5	96.0

Simulation Results

		β_1	β_2	β_3
	TRUE	-1.0	-0.5	2.0
Oracle	AVE	-1.0020	-0.5077	2.0302
	STD	0.2265	0.1870	0.3723
	$\widehat{\text{STD}}$	0.2264	0.1882	0.3670
	CI	0.9420	0.9560	0.9490
Efficient	AVE	-0.9939	-0.5056	2.0201
	STD	0.2610	0.2043	0.4199
	$\widehat{\text{STD}}$	0.2314	0.1952	0.3795
	CI	0.9470	0.9560	0.9450
Locally	AVE	-1.0180	-0.5165	2.0850
Efficient	STD	0.3500	0.3154	0.6732
	$\widehat{\text{STD}}$	0.3903	0.3313	0.6601
	CI	0.9490	0.9620	0.9400

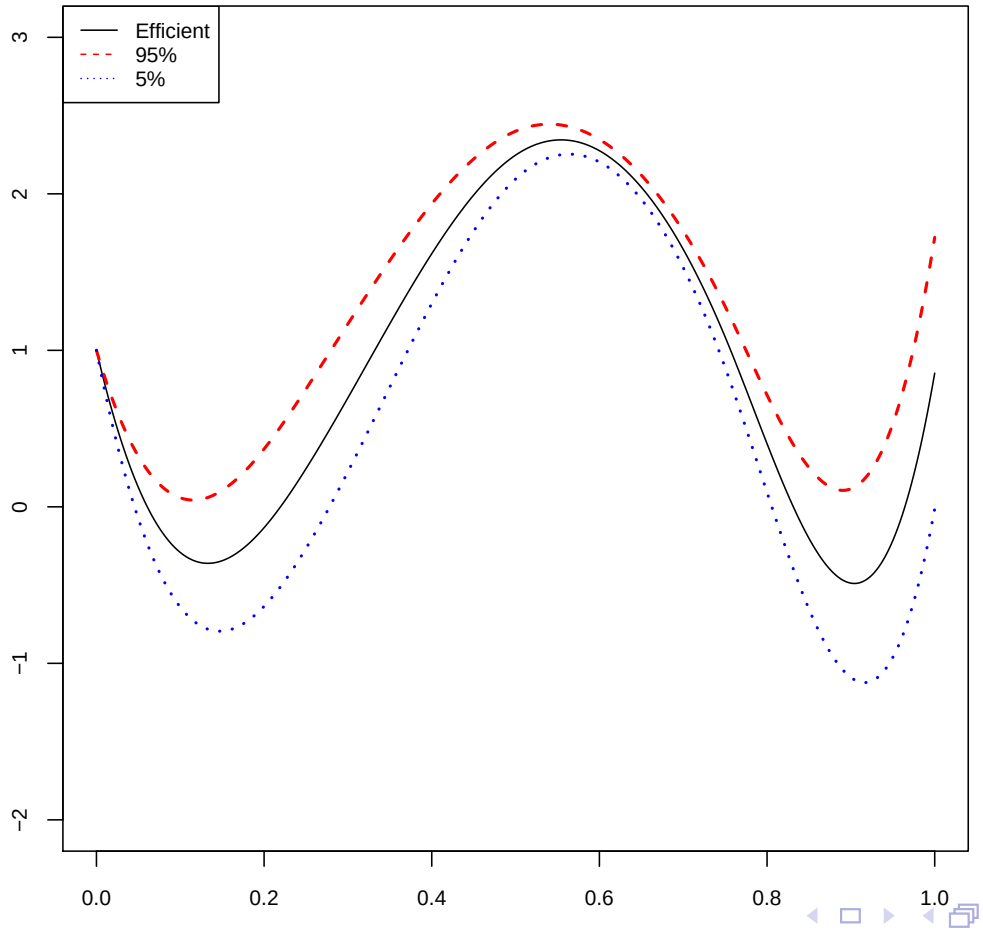
Estimation of $\widehat{\alpha}(t)$ in Simulations





Pollution and CVD Death

	$\widehat{\beta}_1(CO)$	$\widehat{\beta}_2(NO_2)$	$\widehat{\beta}_3(O_3)$	$\widehat{\beta}_4(SO_2)$
Coefficients	-0.2857	-0.9706	-1.8328	1
Standard Errors	0.0797	0.0058	0.0019	0
p-values	0.0003	0.0000	0.0000	0



Very high dimension reduction

eQTL analysis

- Focus: Gene expression quantitative trait loci (eQTL) analysis
  - Key to understand how genetic variations function at molecular level
  - Most prominent way to discover the gene regulation network: how genetic variations govern gene expression
  - Single Nucleotide Polymorphisms (SNPs): common genetic variations
- Goal: Identify association between eQTL SNPs and gene expression level
- GTEx data
  - $n = 119$: observations from lung tissues
  - Y: gene expression level at the gene ENSG00000163191.5
  - X: $p = 322$ (303 SNPs + other covariates)

Traditional method

- Study Y and each SNP.
- The smallest p-value is $0.03 > 0.05/322$, the Bonferroni adjusted p-value.
- No significance is detected.
- No previous study identified any significant SNPs in lung
- Would like to include all the SNPs into the model
- Would like to take inter-SNP correlations into account
- Difficulty: $p > n$

Solution: combine with factor model

- Factor model:

$$X_i = \mathbf{B}\mathbf{h}_i + u_i$$

- Dimension reduction model:

$$f_{Y|\mathbf{h}}(Y_i, \mathbf{h}_i) = f\left(Y_i, \beta^{\mathrm{T}}\mathbf{h}_i\right),$$

- $X_i, u_i : p \times 1$
  $\mathbf{B} : p \times q$
  $\mathbf{h}_i : q \times 1$
  $\beta : q \times d$
  $d < q < p$
- More restricted model in Fan, Xue and Yao (2016)

Estimation

- Factor model
  - Estimate $\mathbf{h}$ through minimizing $\|X - \mathbf{B}\mathbf{H}^{\mathrm{T}}\|_F$ subject to $n^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{H} = \mathbf{I}_q$, $\mathbf{B}^{\mathrm{T}}\mathbf{B}$ is diagonal. $X = (X_1, \ldots, X_n)$, $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_n)^{\mathrm{T}}$.
  - We do not need normality assumptions on $\mathbf{h}$.
- Dimension reduction model
  - We aim at estimating β without normality of $\mathbf{h}_i$.
  - $\phi =$ $[\mathbf{g}(Y_i, \beta^{\mathrm{T}}\mathbf{h}_i) - E\{\mathbf{g}(Y_i, \beta^{\mathrm{T}}\mathbf{h}_i)|\beta^{\mathrm{T}}\mathbf{h}_i\}][\mathbf{a}(\mathbf{h}_i) - E\{\mathbf{a}(\mathbf{h}_i)|\beta^{\mathrm{T}}\mathbf{h}_i\}]$
  - Different $\mathbf{g}, \mathbf{a}$ give: **semi-SIR, semi-SAVE, semi-DR, semi-PHD** (Ma and Zhu 2012)
  - Solve $\sum_{i=1}^{n} \phi(\widehat{\mathbf{h}}_i, \beta) = 0$.

Theoretic properties

- The identifiability of $X_i = \mathbf{B}\mathbf{h}_i + u_i$
  - Issue: $X_i = \mathbf{B}\mathbf{h}_i + \mathbf{u}_i <=> X_i = \mathbf{B}\mathbf{R}\mathbf{R}^{-1}\mathbf{h}_i + \mathbf{u}_i$
  - Added condition: $n^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{H} = \mathbf{I}_q$, $\mathbf{B}^{\mathrm{T}}\mathbf{B}$ is diagonal
  - identifiability relies on uniqueness of svd and $\|p^{-1}\mathbf{B}\mathbf{B}^{\mathrm{T}} - p^{-1}\Sigma_x\|_1 = o_p(1)$
- $\sqrt{n}(\widehat{\beta} - \beta)$ is asymptotic normal
  - Use $\widehat{\mathbf{h}}$ instead of $\mathbf{h}$
  - $p, n \to \infty$, $n^{1/2}p^{-1} \to 0$.
  - Double robustness

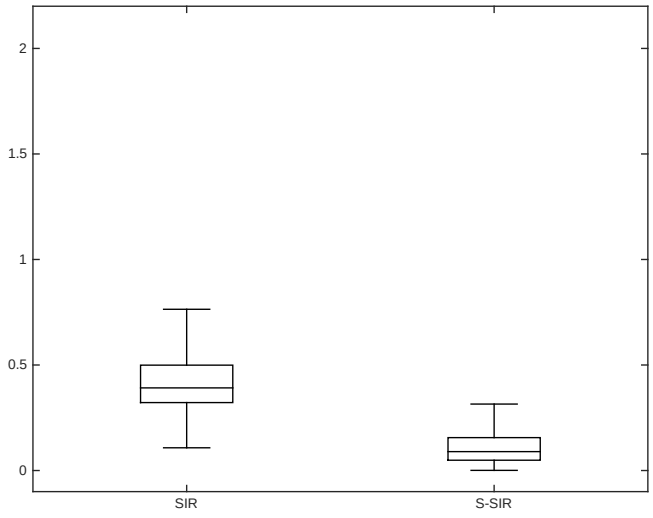
Results for semi-DR

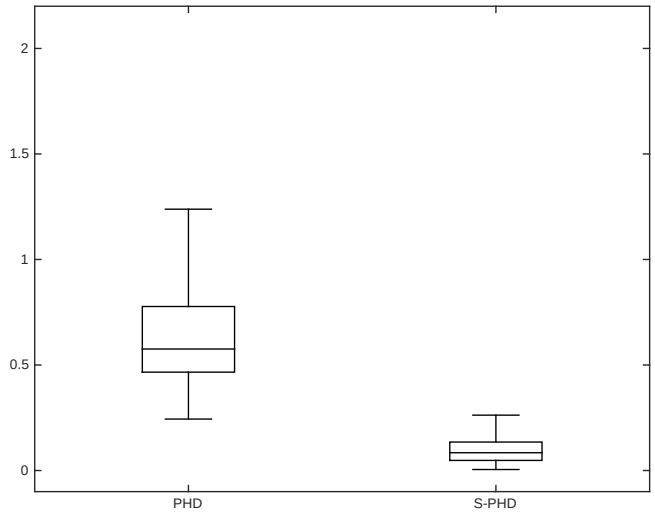
$n = 300, p = 50, q = 6, d = 2$, 1000 simulations

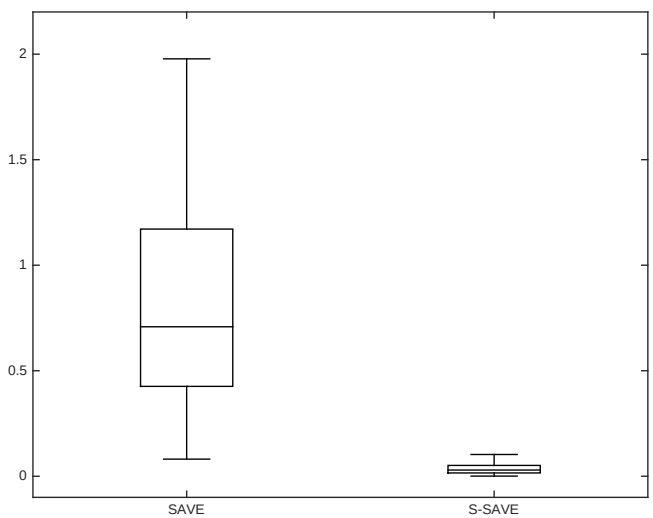
|bias|	std	95% CP
0.0863, 0.0184	0.1256, 0.1433	0.939, 0.934
0.0155, 0.0113	0.1421, 0.1493	0.926, 0.920
0.0667, 0.0492	0.0832, 0.1186	0.949, 0.951
0.0124, 0.0401	0.1185, 0.1481	0.933, 0.951

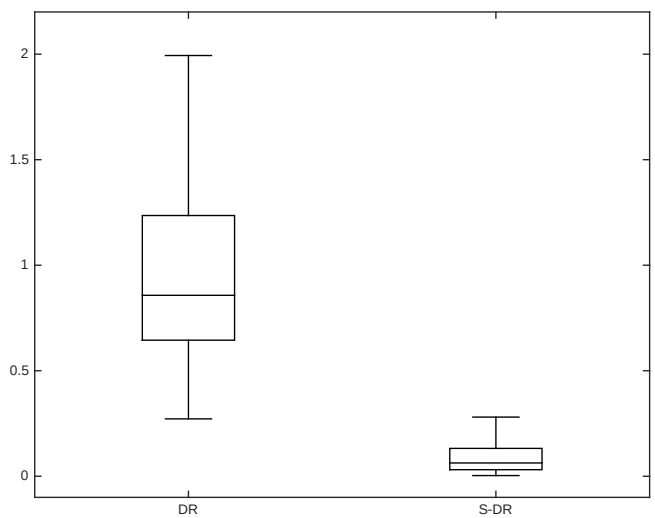
- small biases and variations
- CPs close to the nominal level

Boxplot comparisons of $\|\widehat{\beta} - \beta\|_2$









GTEx data analysis

$n = 119, p = 322$
Determine q dimensional $\widehat{\mathbf{h}}_i$:

- The eigenvectors corresponding to q largest eigenvalues of $X^{\mathrm{T}}X$
- $q = 6$ explains 80% total variation.

Estimate $q \times d$ dimensional $\widehat{\beta}$:

- $\sum_{i=1}^{n} \left[\mathbf{g}(Y_i, \beta^{\mathrm{T}}\widehat{\mathbf{h}}_i) - E\left\{\mathbf{g}(Y_i, \beta^{\mathrm{T}}\widehat{\mathbf{h}}_i) \mid \beta^{\mathrm{T}}\widehat{\mathbf{h}}_i\right\}\right]$
  $\times \left[\mathbf{a}(\widehat{\mathbf{h}}_i) - E\{\mathbf{a}(\widehat{\mathbf{h}}_i) \mid \beta^{\mathrm{T}}\widehat{\mathbf{h}}_i\}\right] = \mathbf{0}$
- $d = 1$ via VIC (Ma and Zhang, 2015)

Cross validation mean predictive errors

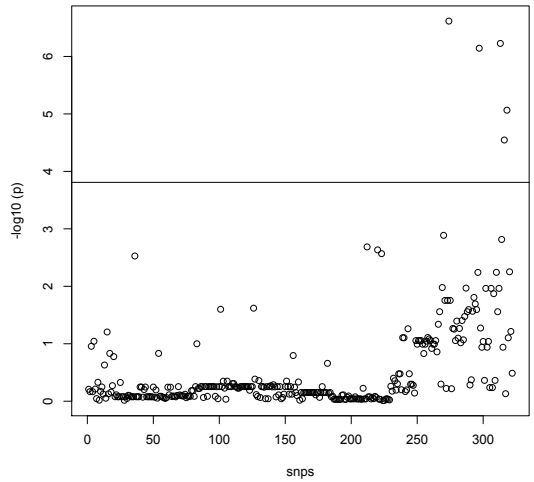
semi-SIR	semi-SAVE	semi-DR	semi-PHD
1.1271	1.1705	1.1018	1.1419
SIR	SAVE	DR	PHD
1.3301	1.2329	1.2824	1.2782

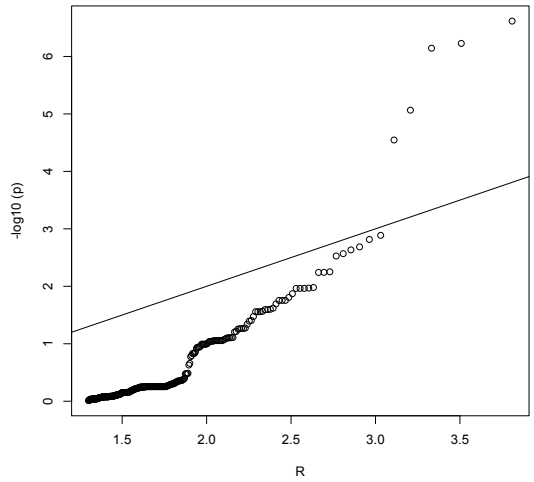
The semiparametric methods outperform the classical methods with smaller mean predictive errors.

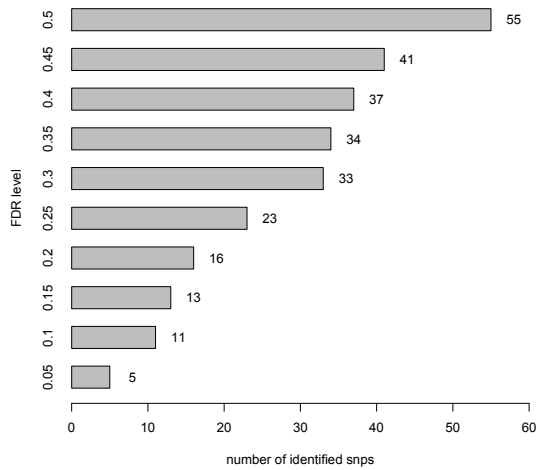
Significant SNPs

$$X_i = \mathbf{B}\mathbf{h}_i + \mathbf{u_i} \longrightarrow \mathbf{h}_i = (\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}X_i - (\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{u_i}$$

$\alpha = \mathbf{B}(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\beta$. Which components in α are nonzero?







Features and Advantages

Method:

- combine dimension reduction and factor model
- handle $p > n$.
- avoid linearity and constant variance condition
- semiparametric approaches are more efficient than classical methods.

Application:

- Include all the SNPs into the model and take inter-SNP correlations into account
- FDR $= 0.05$, identify 5 top SNPs
- Researchers can select sufficient number of SNPs by increase the FDR level.