Semiparametric estimation in high-dimensions

Mladen Kolar (mkolar@chicagobooth.edu)

Collaborators

Sen Na (U Chicago)

Zhaoran Wang (Northwestern)

Zhuoran Yang (Princeton)

Sen Na, Zhuoran Yang, Zhaoran Wang, Mladen Kolar
High-dimensional Varying Index Coefficient Models via Stein's Identity
JMLR (2019), arXiv:1810.07128.

Sen Na, Mladen Kolar
High-dimensional Index Volatility Models via Stein's Identity
Accepted Bernoulli, arXiv:1811.10790

Motivation

Semiparametric models provide flexible fitting of data, while retaining interpretability.

Fitting parameters can be challening.

Our aim is to provide accurate estimation procedures that are efficiently computable and robust.

Outline

1. Varying index coefficient model

2. Index volatilty model

High-dimensional varying index coefficient model

Model (Ma and Song 2015)

$$y = \sum_{j=1}^{d_2} z_j \cdot f_j(\langle \boldsymbol{x}, \boldsymbol{\beta}_j^\star \rangle) + \epsilon$$

- y is the response variable
- $\boldsymbol{x} = (x_1, \ldots, x_{d_1})^\top \in \mathbb{R}^{d_1}$ and $\boldsymbol{z} = (z_1, \ldots, z_{d_2})^\top \in \mathbb{R}^{d_2}$ are given covariates
- ϵ is random noise with $\mathbb{E}[\epsilon \mid \boldsymbol{x}, \boldsymbol{z}] = 0$
- $\boldsymbol{\beta}_j^\star = (\beta_{j1}^\star, \ldots, \beta_{jd_1}^\star)^\top$, $j \in [d_2]$, are the coefficient vectors, which vary with different covariates z_j
- $f_j(\cdot)$ are unknown nonparametric link functions

Identifiability

$$\boldsymbol{\beta}_j^\star \in \{\boldsymbol{\beta} \in \mathbb{R}^{d_1} : \|\boldsymbol{\beta}\|_2 = 1 \text{ and } \beta_1 > 0\}, \quad j = 1, \ldots, d_2.$$

Flexible generalization

Additive single-index model (Chen 1991; Carroll et al. 1997)

- when $z_j = 1$, $j \in [d_2]$
- can be viewed as a two-layer neural network with d_2 hidden nodes

Varying coefficient model (Cleveland, Grosse, and Shyu 1991; Hastie and Tibshirani 1993)

- when $d_1 = 1$ and $\beta_j^\star = 1$, $j = 1, \ldots, d_2$
- wide applications in scientific areas such as economics and medical science (Fan and Zhang 2008)

Varying coefficient models allow the coefficients of $\boldsymbol{z}$ to be smooth functions of $\boldsymbol{x}$, thus incorporating nonlinear interactions between $\boldsymbol{x}$ and $\boldsymbol{z}$.

- easily interpreted in real applications because it inherits features from both single-index model and varying coefficient model
- captures complex multivariate nonlinear structure

Existing estimation approaches

Existing procedures estimate the unknown functions and coefficients iteratively

- with the signal parameters $\{\beta_j^\star\}_{j\in[d_2]}$ fixed, estimate the functions $\{f_j(\cdot)\}_{j\in[d_2]}$ using local polynomial estimator
- with the estimated link functions fixed, one re-estimates the coefficients

The global minimizer has desirable properties (Xue and Wang 2012; Ma and Song 2015)

- the loss function is usually nonconvex
- computationally intractable to obtain the global optima

Question: *Is it possible to estimate signal parameters* $\{\beta_j^\star\}_{j\in[d_2]}$ *with both statistical accuracy and computational efficiency?*

Estimation via the Generalized Stein's Identity

Theorem (Stein et al. 2004):

$\boldsymbol{v} \in \mathbb{R}^d$ is a random vector with differentiable positive density $p_{\boldsymbol{v}} : \mathbb{R}^d \to \mathbb{R}$.

The score function as $S_{\boldsymbol{v}} : \mathbb{R}^d \to \mathbb{R}^d$:

$$S_{\boldsymbol{v}}(\mathbf{v}) = -\nabla \log p_{\boldsymbol{v}}(\mathbf{v})$$

Regularity conditions:

- $|p_{\boldsymbol{v}}(\mathbf{v})| \to 0$ as $\|\mathbf{v}\| \to \infty$
- $\mathbb{E}[|f(\boldsymbol{v})S_{\boldsymbol{v}}(\boldsymbol{v})|] \vee \mathbb{E}[|\nabla f(\boldsymbol{v})|] < \infty$

Then

$$\mathbb{E}[f(\boldsymbol{v})S_{\boldsymbol{v}}(\boldsymbol{v})] = \mathbb{E}[\nabla f(\boldsymbol{v})].$$

When $\boldsymbol{v} \sim N(\mathbf{0}, \boldsymbol{I}_d)$, we have

$$\mathbb{E}[g(\boldsymbol{v})\boldsymbol{v}] = \mathbb{E}[\nabla g(\boldsymbol{v})].$$

Estimation via the Generalized Stein's Identity

Model $y = \sum_{j=1}^{d_2} z_j \cdot f_j(\langle \boldsymbol{x}, \boldsymbol{\beta}_j^\star \rangle) + \epsilon$

Regularity conditions

- $\boldsymbol{x}$, $\boldsymbol{z}$ are independent, the density function $p(\cdot)$ of $\boldsymbol{x}$ is positive and differentiable
- For any $j \in [d_2]$, $\widetilde{f}_j(\mathbf{x}) = f_j(\langle \mathbf{x}, \boldsymbol{\beta}_j^\star \rangle)$ satisfies the conditions of Stein's theorem
- $\mu_j := \mathbb{E}[f_j'(\langle \boldsymbol{x}, \boldsymbol{\beta}_j^\star \rangle)] \neq 0$
- $\boldsymbol{z}$ are standardized with $\mathbb{E}[z_j] = 0$ and $\mathbb{E}[z_j^2] = 1$, $\forall j \in [d_2]$

Under regularity conditions, Stein's identity allows us to extract the unknown coefficient parameter

$$\mathbb{E}[f_j(\langle \boldsymbol{x}, \boldsymbol{\beta}_j^\star \rangle) S(\boldsymbol{x})] = \mathbb{E}[f_j'(\langle \boldsymbol{x}, \boldsymbol{\beta}_j^\star \rangle)] \boldsymbol{\beta}_j^\star = \mu_j \boldsymbol{\beta}_j^\star := \widetilde{\boldsymbol{\beta}}_j$$

Warm-up example

Suppose $\boldsymbol{x} \sim N(\boldsymbol{0}, \boldsymbol{I}_{d_1})$, $\boldsymbol{z} \sim N(\boldsymbol{0}, \boldsymbol{I}_{d_2})$ and $\boldsymbol{x}$ and $\boldsymbol{z}$ are independent.

Stein's identity gives us

$$\mathbb{E}[y \cdot z_k \cdot \boldsymbol{x}] = \sum_{j=1}^{d_2} \mathbb{E}[z_j z_k f_j(\langle \beta_j^\star, \boldsymbol{x} \rangle)\boldsymbol{x}] = \mathbb{E}[f_k(\langle \beta_k^\star, \boldsymbol{x} \rangle)\boldsymbol{x}] = \mu_k \beta_k^\star = \widetilde{\beta}_k$$

The above equation allows us define the following population loss

$$\widetilde{\beta}_k = \arg\min_{\beta_k} L_k(\beta_k) = \arg\min_{\beta_k} \left\{ \|\beta_k\|_2^2 - 2\mathbb{E}[y \cdot z_k \cdot \langle \beta_k, \boldsymbol{x} \rangle] \right\}$$

Warm-up example

Given n i.i.d. samples $\{y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i\}_{i=1}^n$, we estimate $\widetilde{\beta}_k$ as

$$\begin{aligned}
\widehat{\beta}_k &= \arg\min_{\beta_k} \widehat{L}_k(\beta_k) + R_k(\beta_k) \\
&= \arg\min_{\beta_k} \left\{ \|\beta_k\|^2 - \frac{2}{n} \sum_{i=1}^n y_i Z_{ik} \langle \beta_k, \boldsymbol{X}_i \rangle + \lambda_k \|\beta_k\|_1 \right\} \\
&= \mathcal{T}_{\lambda_k/2} \left(\frac{1}{n} \sum_{i=1}^n y_i Z_{ik} \boldsymbol{X}_i \right)
\end{aligned}$$

$\mathcal{T}_\lambda(\boldsymbol{a}) \in \mathbb{R}^d$ is the soft-thresholding function with

$$[\mathcal{T}_\lambda(\boldsymbol{a})]_i = (1 - \lambda/|a_i|)_+ a_i$$

Warm-up example

Theorem:

Suppose regularity conditions hold. Suppose $\|\beta_k^\star\|_0 \leqslant s$, $k \in [d_2]$, $\boldsymbol{x} \sim N(\boldsymbol{0}, \boldsymbol{I}_{d_1})$, components of $\boldsymbol{z}$ are independent with $\|z_k\|_{\psi_2} = \Upsilon_{z_k} \leqslant \Upsilon_{\boldsymbol{z}}$, $k \in [d_2]$, and independent of $\boldsymbol{x}$, and y is sub-exponential with $\|y\|_{\psi_1} \leqslant \Upsilon_y$. If $\lambda_k = 4\Upsilon\sqrt{\log n/n}$, then

$$\|\widehat{\beta}_k - \widetilde{\beta}_k\|_2 \leqslant \frac{3}{2}\sqrt{s}\lambda_k \quad \text{and} \quad \|\widehat{\beta}_k - \widetilde{\beta}_k\|_1 \leqslant 6s\lambda_k, \quad \forall k \in [d_2]$$

with probability at least $1 - d_2 d_1/n^2$.

For centered random variable x,

- $\|x\|_{\psi_1} = \sup_{p \geqslant 1} p^{-1}(\mathbb{E}|x|^p)^{1/p}$;
- $\|x\|_{\psi_2} = \sup_{p \geqslant 1} p^{-1/2}(\mathbb{E}|x|^p)^{1/p}$.

We call x a sub-exponential random variable if $\|x\|_{\psi_1} < \infty$.
We call x a sub-Gaussian random variable if $\|x\|_{\psi_2} < \infty$.

General Setup

We will assume that there exists a constant $M_p > 0$ such that

$$\mathbb{E}[y^p] \vee \mathbb{E}[S(\boldsymbol{x})_j^p] \vee \mathbb{E}[z_k^p] \leqslant M_p, \quad \forall j \in [d_1], k \in [d_2].$$

- p is either 4 or 6

Study two settings

- estimation under sparsity
- estimation under low-rank assumption

Sparse Vector Recovery

When $\boldsymbol{x}$ is not Gaussian and $\mathbb{E}[z_j z_k] = 0$ for $j \neq k$

$$\mathbb{E}[y \cdot z_k \cdot S(\boldsymbol{x})] = \sum_{j=1}^{d_2} \mathbb{E}[z_j z_k f_j(\langle \beta_j^\star, \boldsymbol{x} \rangle) S(\boldsymbol{x})] = \mathbb{E}[f_k(\langle \beta_k^\star, \boldsymbol{x} \rangle) S(\boldsymbol{x})] = \mu_k \beta_k^\star = \widetilde{\beta}_k$$

$$\begin{aligned}
\widehat{\beta}_k &= \arg\min_{\beta_k} \bar{L}_k(\beta_k) + R_k(\beta_k) \\
&= \arg\min_{\beta_k} \left\{ \|\beta_k\|^2 - \frac{2}{n} \sum_{i=1}^{n} \check{y}_i \widetilde{Z_{ik}} \langle \beta_k, \widetilde{S(\boldsymbol{X}_i)} \rangle + \lambda_k \|\beta_k\|_1 \right\} \\
&= \mathcal{T}_{\lambda_k/2} \left(n^{-1} \sum_{i=1}^{n} \check{y}_i \widetilde{Z_{ik}} \widetilde{S(\boldsymbol{X}_i)} \right)
\end{aligned}$$

Given a threshold $\tau > 0$, the hard truncation

$$\check{\boldsymbol{v}} \in \mathbb{R}^d \qquad [\check{\boldsymbol{v}}]_i = \begin{cases} v_i & \text{if } |v_i| \leqslant \tau \\ 0 & \text{o/w.} \end{cases}$$

Sparse Vector Recovery

Theorem:

Suppose regularity conditions hold.
Suppose $\|\beta_k^\star\|_0 \leqslant s$, $\forall k \in [d_2]$, and $\mathbb{E}[z_j z_k] = 0$ for $j \neq k$.
If $\lambda_k = 76\sqrt{M_6 \log d_1 d_2 / n}$ and $\tau = (M_6 n / \log d_1 d_2)^{1/6}/2$, then

$$\|\widehat{\beta}_k - \widetilde{\beta}_k\|_2 \leqslant \frac{3}{2}\sqrt{s}\lambda_k \quad \text{and} \quad \|\widehat{\beta}_k - \widetilde{\beta}_k\|_1 \leqslant 6s\lambda_k, \quad \forall k \in [d_2],$$

with probability at least $1 - 2/d_1^2 d_2^2$.

In particular, with high probability

$$\|\widehat{\beta}_k - \widetilde{\beta}_k\|_2 \lesssim \sqrt{\frac{s \log d_1 d_2}{n}} \quad \text{and} \quad \|\widehat{\beta}_k - \widetilde{\beta}_k\|_1 \lesssim s\sqrt{\frac{\log d_1 d_2}{n}}.$$

The proof relies on establishing that

$$P\left(\|\nabla \bar{L}_k(\widetilde{\beta}_k)\|_\infty \leqslant 38\sqrt{\frac{M_6 \log d_1 d_2}{n}}, \quad \forall k \in [d_2] \right) \geqslant 1 - \frac{2}{d_1^2 d_2^2}.$$

Sparse Recovery

Relaxing the condition $\mathbb{E}[z_j z_k] = 0$

Let $\boldsymbol{\Sigma}^\star = \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top] \in \mathbb{R}^{d_2 \times d_2}$ and $\boldsymbol{\Omega}^\star = (\boldsymbol{\Sigma}^\star)^{-1}$ with

$$\mathcal{F}_w^K = \left\{ \boldsymbol{\Omega} \succeq \mathbf{0} : \|\boldsymbol{\Omega}\|_{0,\infty} \leqslant w, \|\boldsymbol{\Omega}\|_2 \leqslant K, \|\boldsymbol{\Omega}^{-1}\|_2 \leqslant K \right\}$$

The identifiability relationship

$$\begin{aligned} \mathbb{E}[y \cdot S(\boldsymbol{x})\boldsymbol{z}^T]\boldsymbol{\Omega}^\star &= \sum_{j=1}^{d_2} \mathbb{E}[f_j(\langle \beta_j^\star, \boldsymbol{x} \rangle) S(\boldsymbol{x})] \mathbb{E}[z_j \cdot \boldsymbol{z}^T] \boldsymbol{\Omega}^\star \\ &= \sum_{j=1}^{d_2} \widetilde{\beta}_j \boldsymbol{e}_j^T \boldsymbol{\Sigma}^\star \boldsymbol{\Omega}^\star = (\widetilde{\beta}_1, ..., \widetilde{\beta}_{d_2}) = \widetilde{\boldsymbol{B}}, \end{aligned}$$

where $\boldsymbol{e}_j \in \mathbb{R}^{d_2}$ is the j-th canonical basis vector

Sparse Recovery

Our estimator

$$\widehat{\boldsymbol{B}} = \arg\min_{\boldsymbol{B}} \left\{ \|\boldsymbol{B}\|_F^2 - \frac{2}{n} \sum_{i=1}^{n} \langle \breve{y}_i \cdot \widetilde{S(\boldsymbol{X}_i)} \breve{\boldsymbol{Z}}_i^T \widehat{\boldsymbol{\Omega}}, \boldsymbol{B} \rangle + \lambda \|\boldsymbol{B}\|_{1,1} \right\}$$

Where $\widehat{\boldsymbol{\Omega}}$ is obtained using the CLIME (Cai, Liu, and Luo 2011)

$$\begin{aligned} &\min \ \|\boldsymbol{\Omega}\|_{1,1}, \\ &\text{s.t.} \ \ \|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Omega} - \boldsymbol{I}_{d_2}\|_{\max} \leqslant \gamma \end{aligned}$$

with

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} \breve{\boldsymbol{Z}}_i \breve{\boldsymbol{Z}}_i^T \tag{1}$$

Sparse Recovery

Lemma:
If $\tau = (M_4 n/\log d_2)^{1/4}/2$ and $\gamma = 12\|\mathbf{\Omega}^\star\|_1\sqrt{M_4 \log d_2/n}$, then

$$P\left(\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}^\star\|_2 \leqslant 96\|\mathbf{\Omega}^\star\|_1^2 w\sqrt{M_4 \log d_2/n}\right) \geqslant 1 - \frac{2}{d_2^2},$$

and

$$P\left(\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}^\star\|_{\max} \leqslant 48\|\mathbf{\Omega}^\star\|_1^2\sqrt{M_4 \log d_2/n}\right) \geqslant 1 - \frac{2}{d_2^2}.$$

- γ depends on $\|\mathbf{\Omega}^\star\|_1$, which is unknown
- the dependence could be removed by using a self-calibrated estimator

Lemma:
If $\tau = (M_6 n/\log d_1 d_2)^{1/6}/2$, then

$$\left\|\mathbb{E}[y \cdot S(\boldsymbol{x})\boldsymbol{z}^T] - \frac{1}{n}\sum_{i=1}^n \breve{y}_i \cdot \widetilde{S(\boldsymbol{X}_i)}\breve{\boldsymbol{Z}}_i^T\right\|_{\max} \leqslant 19\sqrt{\frac{M_6 \log d_1 d_2}{n}}$$

with probability at least $1 - 2/d_1^2 d_2^2$.

Sparse Recovery

Theorem:

Suppose regularity conditions hold.
Suppose $\|\beta_k^\star\|_0 \leqslant s$, $k \in [d_2]$ and that the precision matrix estimator $\widehat{\mathbf{\Omega}}$ satisfies

$$P(\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}^\star\|_{\max} \leqslant \widetilde{\mathcal{H}}(n, d_2)) \geqslant 1 - \widetilde{\mathcal{P}}(n, d_2).$$

If $\tau = (M_6 n / \log d_1 d_2)^{1/6}/2$ and

$$\lambda \geqslant 76\|\mathbf{\Omega}^\star\|_1 \sqrt{\frac{M_6 \log d_1 d_2}{n}} + 4 \max_{j \in [d_2]} |\mu_j| \cdot \|\boldsymbol{B}^\star \mathbf{\Sigma}^\star\|_\infty \widetilde{\mathcal{H}}(n, d_2),$$

then

$$\|\widehat{\boldsymbol{B}} - \widetilde{\boldsymbol{B}}\|_F \leqslant 2\sqrt{s d_2}\lambda \quad \text{and} \quad \|\widehat{\boldsymbol{B}} - \widetilde{\boldsymbol{B}}\|_{1,1} \leqslant 8 s d_2 \lambda,$$

with probability at least $1 - 2/d_1^2 d_2^2 - \widetilde{\mathcal{P}}(n, d_2)$.

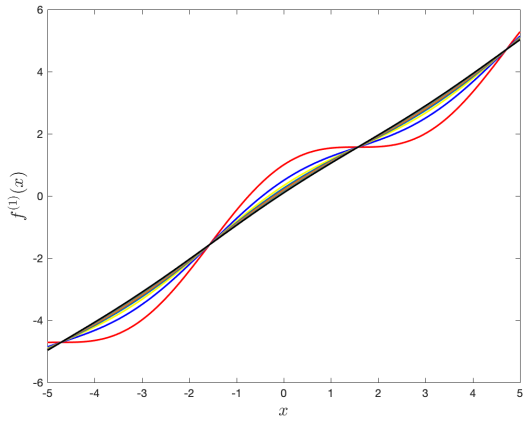
Sparse Recovery

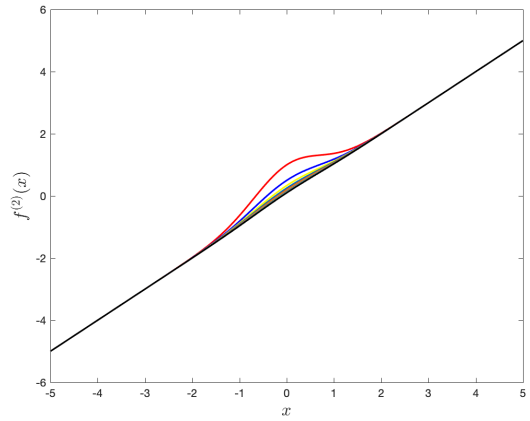
The error in estimating the precision matrix only contributes higher order error terms and

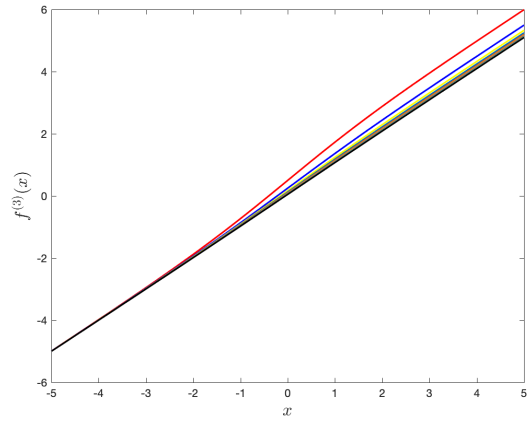
$$\|\widehat{\boldsymbol{B}} - \widetilde{\boldsymbol{B}}\|_F \lesssim \sqrt{\frac{s \log d_1 d_2}{n}} \quad \text{and} \quad \|\widehat{\boldsymbol{B}} - \widetilde{\boldsymbol{B}}\|_{1,1} \lesssim s\sqrt{\frac{\log d_1 d_2}{n}}$$

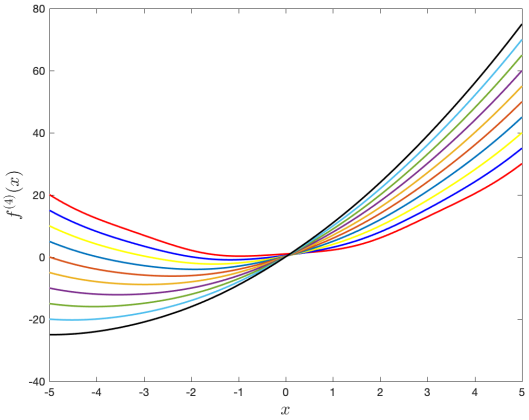
Simulations

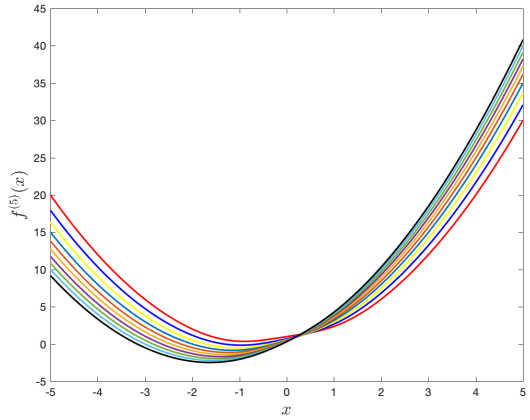
When k varies from 1 to 10, the line moves from red to black.

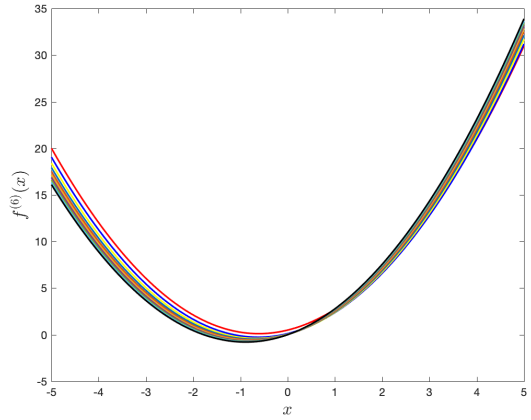












Simulations

Distribution of x

Distribution	parameter	score function
Gaussian	$\mu = 0;\ \sigma = 1$	$s(x) = x$
Beta	$\alpha = 8;\ \beta = 8$	$s(x) = \frac{14x-7}{(1-x)x}$
Gamma	$k = 8;\ \theta = 0.1$	$s(x) = 10 - \frac{7}{x}$
Student's t	$\nu = 13$	$s(x) = \frac{14x}{13+x^2}$
Rayleigh	$\sigma = 1$	$s(x) = x - \frac{1}{x}$
Weibull	$k = 7;\ \lambda = 1$	$s(x) = 7x^6 - \frac{6}{x}$

Simulations — Sparse Recovery

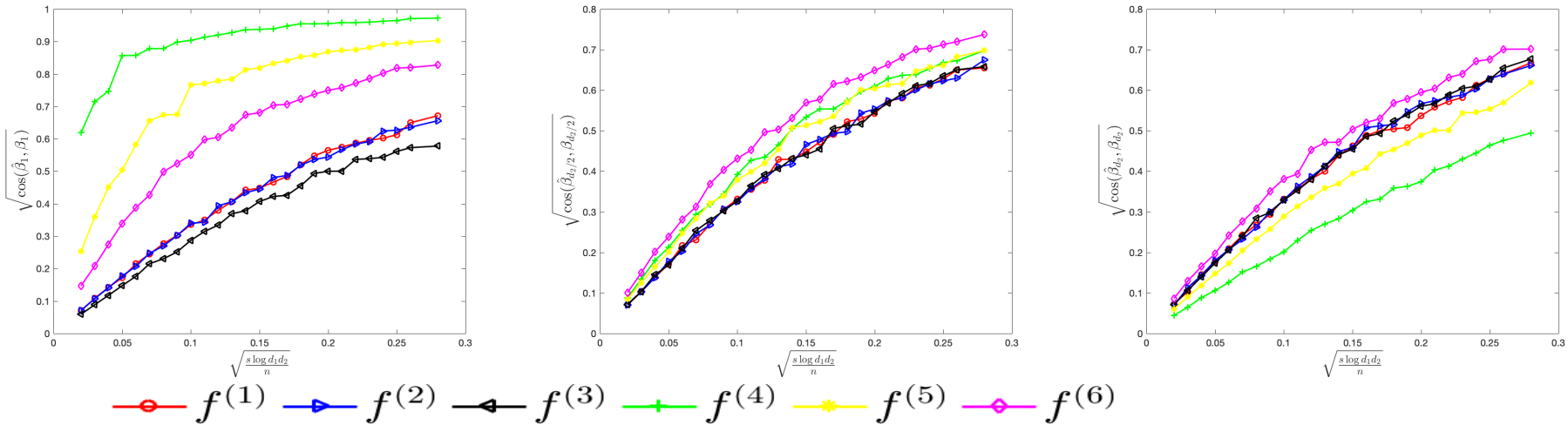
Simulation setting

- fully sparse $\boldsymbol{B}^\star$ with dependent covariate $\boldsymbol{z}$
- set $d_1 = 100$, $d_2 = 50$, $s = 10$, vary n
- $\boldsymbol{X}_i$ is independent and identically generated
- $Z_{ik} \in \{-1, 1\}$ with equal probability and independent from other coordinates
- the support S_k drawn uniformly at random, $\{\beta^\star_{kl}\}_{l \in S_k} \sim \frac{1}{\sqrt{s}} \cdot \mathrm{Unif}(\{-1, 1\})$

- $\lambda_k = 30\sqrt{\log d_1 d_2 / n}$ and $\tau = 2(n/\log d_1 d_2)^{1/6}$

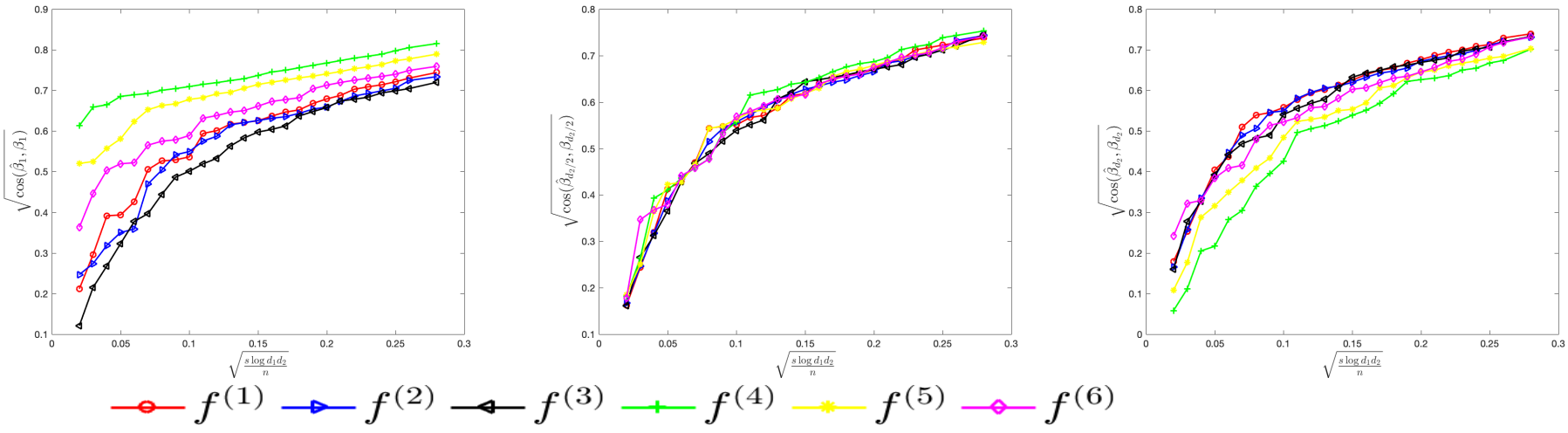
Simulations — Sparse Recovery

Gaussian design



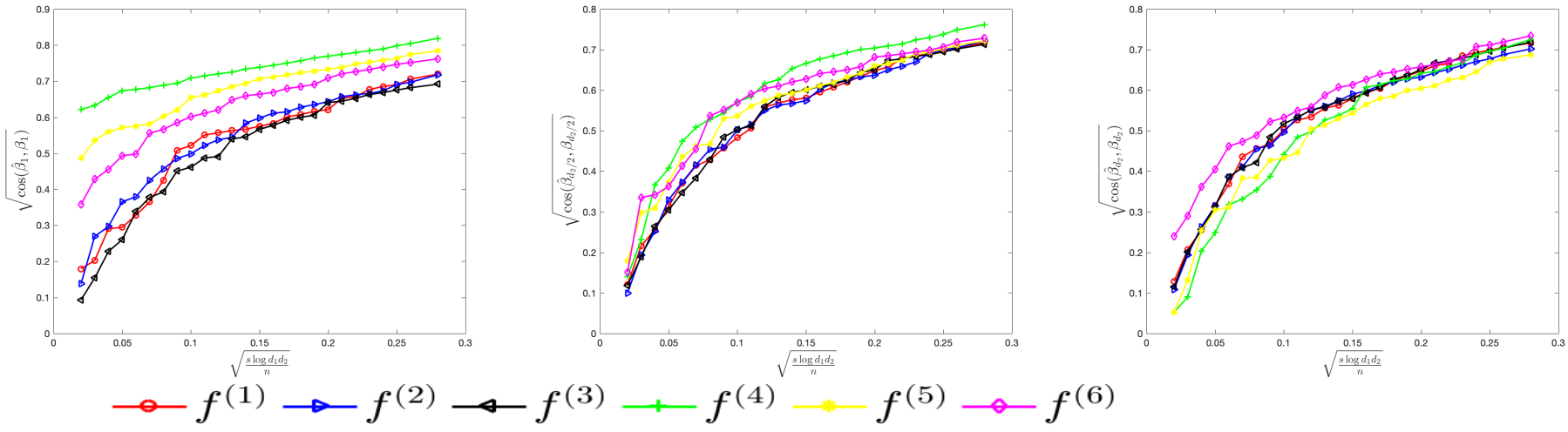
Simulations — Sparse Recovery

Beta design



Simulations — Sparse Recovery

Gamma design



Simulations — Sparse Recovery

Simulation setting

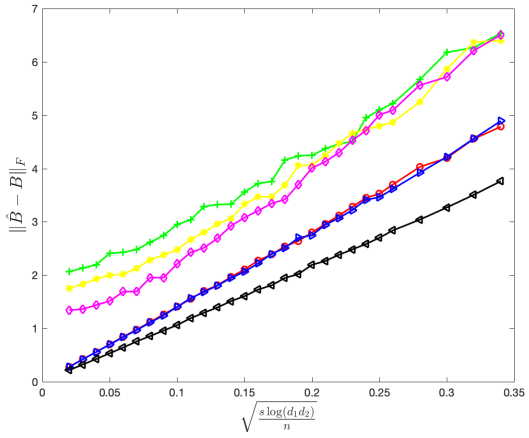
- set $d_1 = 100$, $d_2 = 20$, $s = 10$, vary n
- the support S drawn uniformly at random; each entry on the support is $\{-\frac{1}{\sqrt{s}}, \frac{1}{\sqrt{s}}\}$ with equal probability
- $\boldsymbol{X}_i$ is independent and identically generated
- $\boldsymbol{Z}$ follows a Gaussian copula with the sparse precision matrix $\Theta = (\theta_{ij})_{i,j=1}^{d_2}$

$$\theta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0.2 & \text{if } |i-j| = 1, \\ 0 & \text{otherwise.} \end{cases}$$

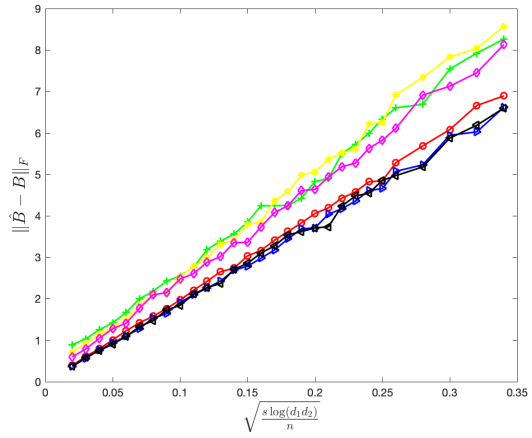
with marginal distribution t_7

- $\tau = 2(n/\log d_1 d_2)^{1/6}$ and $\lambda = 10\sqrt{\log d_1 d_2 / n}$
- for estimation of the precision matrix, we use the truncation threshold $2(n/\log d_2)^{1/4}$ and $\gamma = 10\sqrt{\log d_2 / n}$

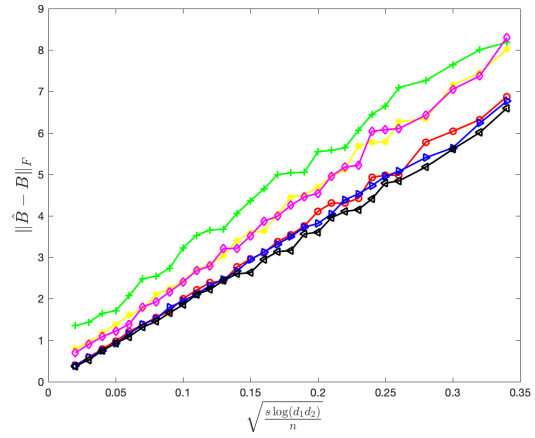
Simulations — Sparse Recovery



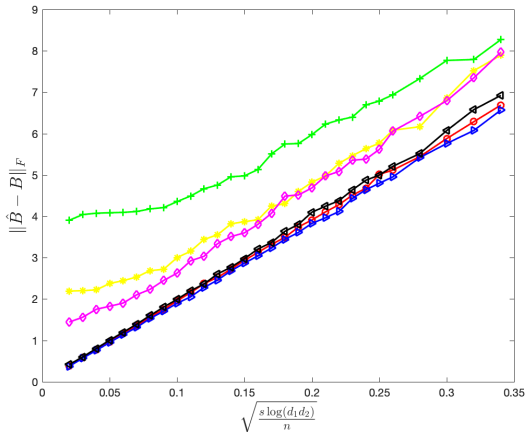
(a) Gaussian



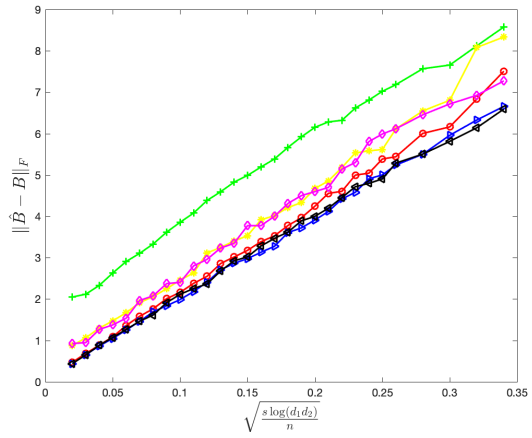
(b) Beta



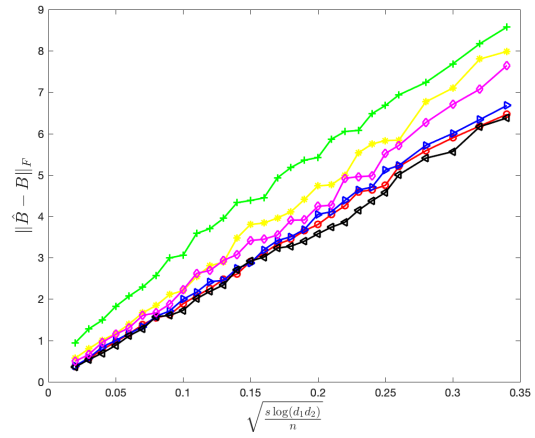
(c) Gamma



(d) t_{13}



(e) Rayleigh



(f) Weibull

$f^{(1)}$ $f^{(2)}$ $f^{(3)}$ $f^{(4)}$ $f^{(5)}$ $f^{(6)}$

Real Data Illustration

Coffea canephora genetic data set — collects three traits (phenotypes) - production of coffee beans (y_1) - leaf rust incidence (y_2) - yield of green beans (y_3)

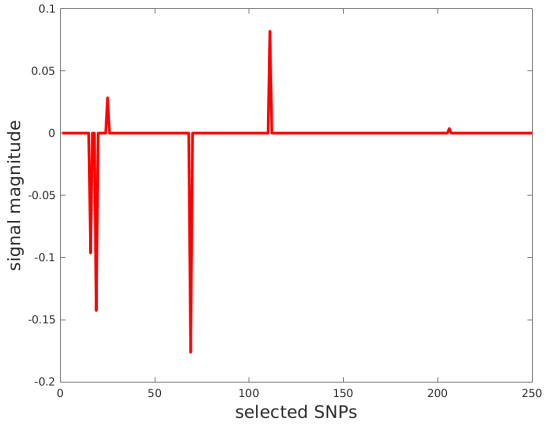
From two recurrent selection populations (x_1) of *Coffea canephora* and each one of traits is evaluated at two locations (x_2).

For each individual sample, the single nucleotide polymorphisms (SNPs)—genotype $\{z_j\}$—are identified by Genotyping-by-Sequencing.

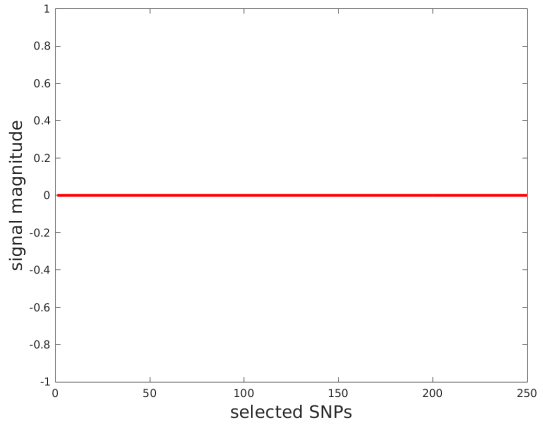
The population group (x_1) and the evaluation location (x_2) are two confounders which may modify the effect of each SNP (z_j) on traits (y_1, y_2, y_3).

$$y_i = \sum_{j=1}^{d_2} z_j \cdot f_j\left((\beta_j^{1,i})^\star x_1 + (\beta_j^{2,i})^\star x_2\right) + \epsilon, \quad \forall i = 1, 2, 3$$

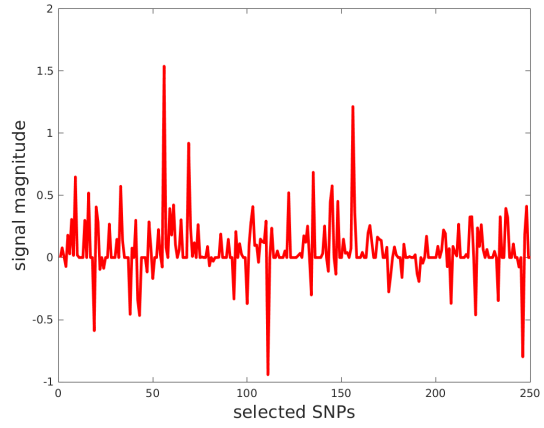
Real Data Illustration



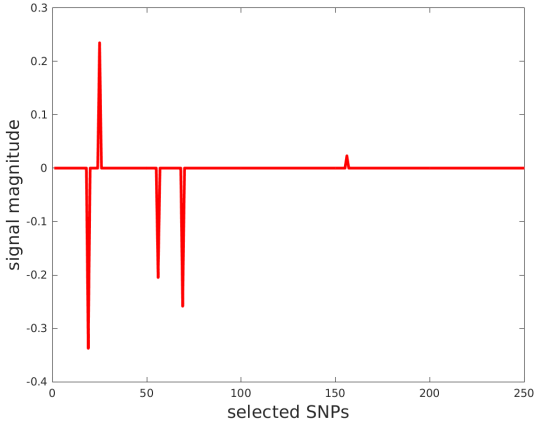
(g) Production



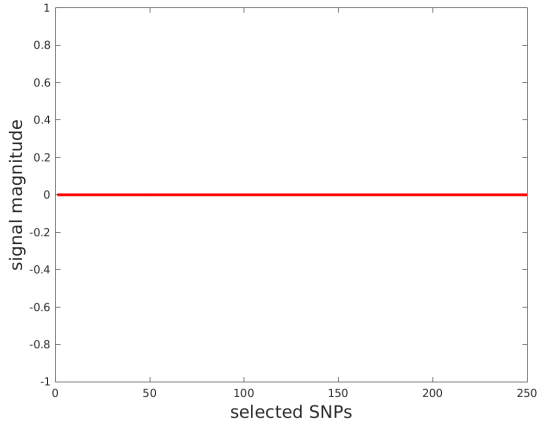
(h) Leaf rust



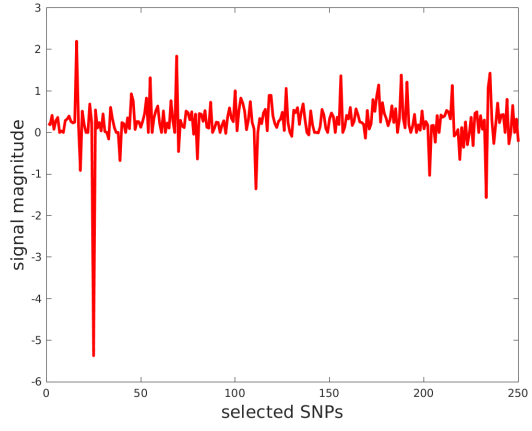
(i) Green beans



(j) Production



(k) Leaf rust



(l) Green beans

High-Dimensional Index Volatility Models

Model

$$y \mid \mathbf{x} = f(\langle \beta^\star, \mathbf{x} \rangle) + g({\boldsymbol{G}^\star}^T \mathbf{x})\epsilon$$

- y is the response variable
- $\mathbf{x} \in \mathbb{R}^d$ is the vector of predictors
- ϵ is random noise with $\mathbb{E}[\epsilon] = 0$ and $\mathbb{E}[\epsilon^2] = 1$, independent of $\mathbf{x}$
- $\beta^\star \in \mathbb{R}^d$ and $\boldsymbol{G}^\star = (\gamma_1^\star, \ldots, \gamma_v^\star) \in \mathbb{R}^{d \times v}$ are unknown parametric components
- $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R}^v \to \mathbb{R}$ are unknown nonparametric link functions

Identifiability

$${\beta^\star}^T \beta^\star = 1 \qquad \text{and} \qquad {\boldsymbol{G}^\star}^T \boldsymbol{G}^\star = I_v$$

Related work

Variance function is constant — homoscedastic single index model

- applications (Sharpe 1963; Collins and Barry 1986; Stock and Watson 1988)
- estimation procedures (Ichimura 1993; Härdle, Hall, and Ichimura 1993; Horowitz and Härdle 1996; Xia et al. 2002; Delecroix, Hristache, and Patilea 2006)
- sliced inverse regression (Li 1991)
- high-dimensional estimation (Babichev and Bach 2018; Plan and Vershynin 2016; Yang, Balasubramanian, and Liu 2017)

Conditional heteroscedasticity

- estimating the function g does not only help in the estimation of the mean, but is interesting in its own right (Box and Hill 1974; Bickel 1978; Box and Meyer 1986)
- when $v = 1$, Härdle, Hall, and Ichimura (1993) first studies single index volatility model (Xia, Tong, and Li 2002; Zhang 2018)
- Chiou and Müller (2004) studies multiple index models with purely nonparametric variance function

Second-order Stein's identity

Theorem (Janzamin, Sedghi, and Anandkumar 2014):

$\mathbf{X} \in \mathbb{R}^d$ is a random vector with differentiable positive density $p_{\mathbf{X}} : \mathbb{R}^d \to \mathbb{R}$.

The second-order score function $H_{\mathbf{X}} : \mathcal{X} \to \mathbb{R}^{d \times d}$ is

$$H_{\mathbf{X}}(\boldsymbol{x}) = \nabla_{\boldsymbol{x}}^2 p_{\mathbf{X}}(\boldsymbol{x}) / p_{\mathbf{X}}(\boldsymbol{x})$$

Regularity conditions:

- first order regularity conditions hold
- $\mathbb{E}[|f(\mathbf{X}) \cdot H_{\mathbf{X}}(\mathbf{X})|] \vee \mathbb{E}[|\nabla_{\boldsymbol{x}}^2 f(\mathbf{X})|] < \infty$

$$\mathbb{E}[f(\mathbf{X}) \cdot H_{\mathbf{X}}(\mathbf{X})] = \mathbb{E}[\nabla_{\boldsymbol{x}}^2 f(\mathbf{X})]$$

When $\mathbf{X} \sim N(\mathbf{0}, \boldsymbol{I}_d)$, then $H(\mathbf{X}) = \mathbf{X}\mathbf{X}^T - I_d$ and we have

$$\mathbb{E}[f(\mathbf{X}) \cdot (\mathbf{X}\mathbf{X}^T - I_d)] = \mathbb{E}[\nabla^2 f(\mathbf{X})].$$

Single Index Volatility Model

Start with the case when $v = 1$

$$y \mid \mathbf{x} = f(\langle \mathbf{x}, \beta^\star \rangle) + g(\langle \mathbf{x}, \gamma^\star \rangle)\epsilon$$

First order estimation

$$\mathbb{E}\big[\big(y - f(\langle \mathbf{x}, \beta^\star \rangle)\big)^2 S(\mathbf{x})\big] = \mathbb{E}[\epsilon^2 g^2(\langle \mathbf{x}, \gamma^\star \rangle) S(\mathbf{x})] = 2\mu_1 \gamma^\star$$

Second order estimation

$$U^\star := \mathbb{E}[(y - f(\langle \mathbf{x}, \beta^\star \rangle))^2 H(\mathbf{x})] = \mathbb{E}[\epsilon^2 g^2(\langle \mathbf{x}, \gamma^\star \rangle) H(\mathbf{x})] = 2\mu_2 \gamma^\star {\gamma^\star}^T$$

Simulations

- Distribution of covariate **x**

Distribution	Parameter	First-order score	Second-order score
Gaussian	$\mu = 0$, $\sigma = 1$	$S(x) = x$	$H(x) = x^2 - 1$
Student's t	degree of freedom 13	$S(x) = \frac{14x}{13+x^2}$	$H(x) = \frac{224x^2}{(13+x^2)^2} - \frac{14}{13+x^2}$
Gamma	$k = 13$, $\theta = 2$	$S(x) = \frac{1}{2} - \frac{12}{x}$	$H(x) = \frac{132}{x^2} - \frac{12}{x} + \frac{1}{4}$

- mean link function t $f(x) = 2x + \cos(x)$
- variance link functions

$$g_1(x) = x + 1 + \cos(x); \quad g_2(x) = x + 1 + \exp(-x^2);$$

$$g_3(x) = x + 1 + \frac{\exp(x)}{(1+\exp(x))^2}; \quad g_4(x) = x^2 + x + \cos(x);$$

$$g_5(x) = x^2 + x + \exp(-x^2); \quad g_6(x) = x^2 + x + \frac{\exp(x)}{(1+\exp(x))^2}$$

- $d = 100$ and $s = 10$

Score estimation

$\{\zeta_j(x)\}_{j=1}^m$ is a set of basis functions; $m = 101$

- $\zeta_j(x) = \frac{1}{\sqrt{2\pi h^2}} \exp(-\frac{(x-z_j)^2}{2h^2})$ be Gaussian kernels with $h = 0.5$ and $z_j = -5 + 0.25(j-1)$

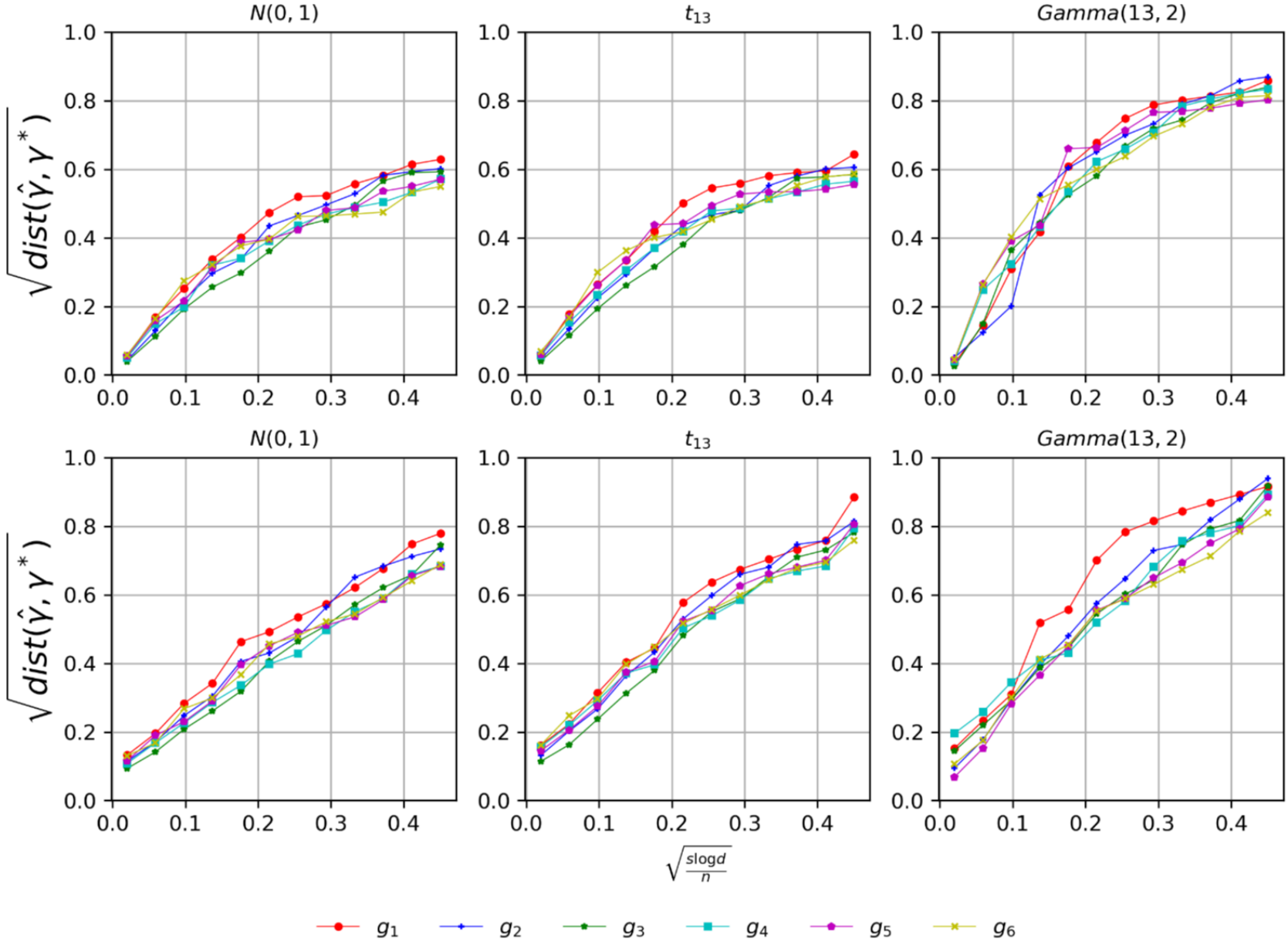
Suppose $S(x) = \sum_{j=1}^m \nu_j^\star \zeta_j(x)$

Score matching estimator (Hyvärinen 2005)

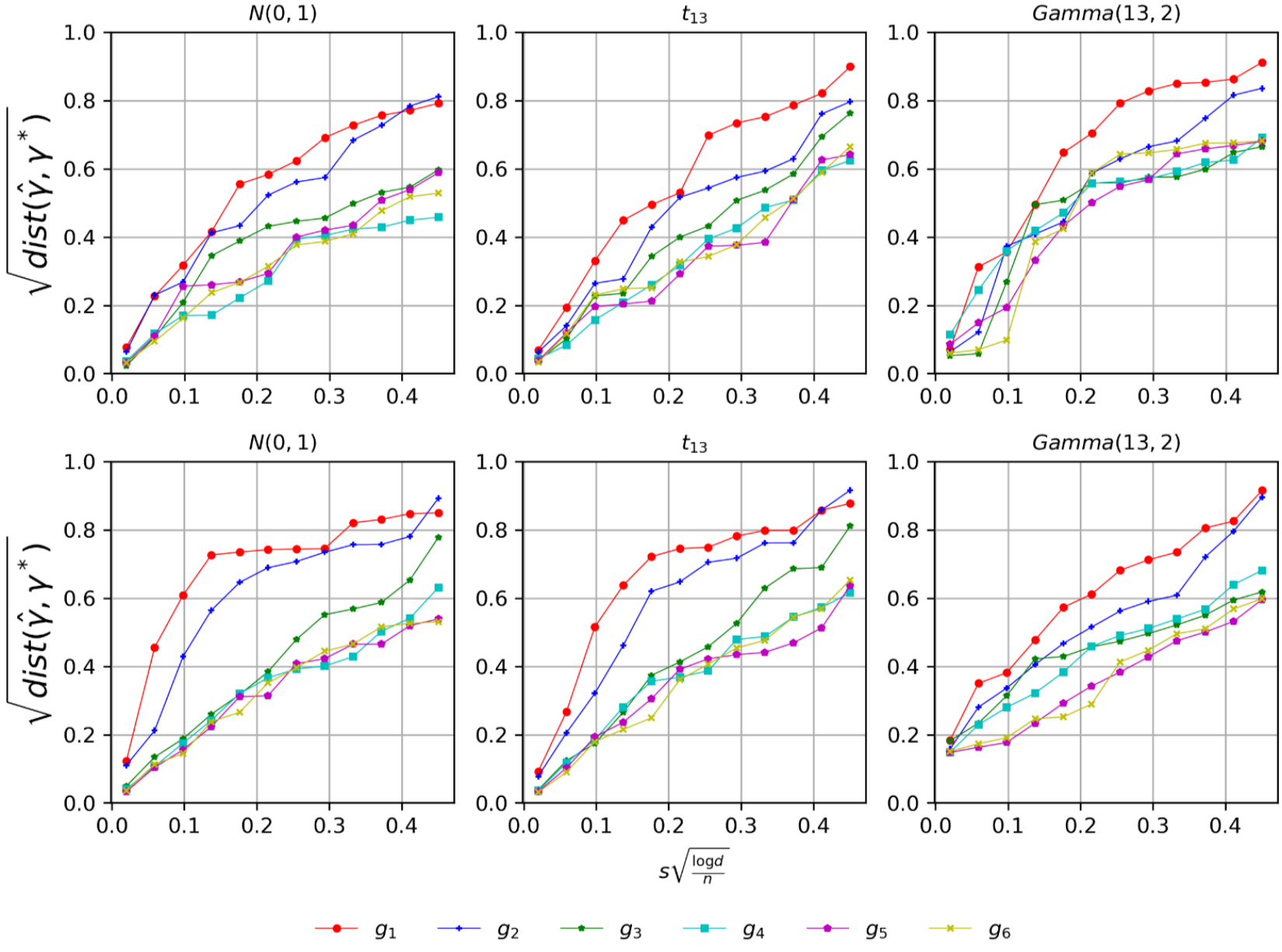
$$\widehat{\boldsymbol{\nu}} = \left(\frac{1}{n} \sum_{i=1}^n \boldsymbol{\zeta}(x_i) \boldsymbol{\zeta}(x_i)^T + \alpha I_m \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \boldsymbol{\zeta}'(x_i) \right)$$

- $\alpha = 0.01$

First order estimation



Second order estimation



Thank you!

References I

Babichev, Dmitry, and Francis Bach. 2018. "Slice Inverse Regression with Score Functions." *Electron. J. Stat.* 12 (1): 1507–43. https://doi.org/10.1214/18-EJS1428.

Bickel, P. J. 1978. "Using Residuals Robustly. I. Tests for Heteroscedasticity, Nonlinearity." *Ann. Statist.* 6 (2): 266–91. http://links.jstor.org/sici?sici=0090-5364(197803)6:2<266:URRITF>2.0.CO;2-Z&origin=MSN.

Box, George E. P., and William J. Hill. 1974. "Correcting Inhomogeneity of Variance with Power Transformation Weighting." *Technometrics* 16: 385–89. https://doi.org/10.2307/1267668.

Box, George E. P., and R. Daniel Meyer. 1986. "An Analysis for Unreplicated Fractional Factorials." *Technometrics* 28 (1): 11–18. https://doi.org/10.2307/1269599.

Cai, T. Tony, W. Liu, and X. Luo. 2011. "A Constrained ℓ_1 Minimization Approach to Sparse Precision Matrix Estimation." *J. Am. Stat. Assoc.* 106 (494): 594–607.

References II

Carroll, R. J., Jianqing Fan, Irène Gijbels, and M. P. Wand. 1997. "Generalized Partially Linear Single-Index Models." *J. Amer. Statist. Assoc.* 92 (438): 477–89. https://doi.org/10.2307/2965697.

Chen, Hung. 1991. "Estimation of a Projection-Pursuit Type Regression Model." *Ann. Statist.* 19 (1): 142–57. https://doi.org/10.1214/aos/1176347974.

Chiou, Jeng-Min, and Hans-Georg Müller. 2004. "Quasi-Likelihood Regression with Multiple Indices and Smooth Link and Variance Functions." *Scand. J. Statist.* 31 (3): 367–86. https://doi.org/10.1111/j.1467-9469.2004.02-117.x.

Cleveland, W. S., E. Grosse, and W. M. Shyu. 1991. "Local Regression Models." In *Statistical Models in S*, edited by J. M. Chambers and Trevor J. Hastie, 309–76.

Collins, Robert A, and Peter J Barry. 1986. "Risk Analysis with Single-Index Portfolio Models: An Application to Farm Planning." *American Journal of Agricultural Economics* 68 (1). Oxford University Press: 152–61.

References III

Delecroix, Michel, Marian Hristache, and Valentin Patilea. 2006. "On Semiparametric *M*-Estimation in Single-Index Regression." *J. Statist. Plann. Inference* 136 (3): 730–69. https://doi.org/10.1016/j.jspi.2004.09.006.

Fan, Jianqing, and Wenyang Zhang. 2008. "Statistical Methods with Varying Coefficient Models." *Statistics and Its Interface* 1 (1). NIH Public Access: 179–95.

Hastie, Trevor J., and Robert J. Tibshirani. 1993. "Varying-Coefficient Models." *J. R. Stat. Soc. B* 55 (4): 757–96.

Härdle, Wolfgang, Peter Hall, and Hidehiko Ichimura. 1993. "Optimal Smoothing in Single-Index Models." *Ann. Statist.* 21 (1): 157–78. https://doi.org/10.1214/aos/1176349020.

Horowitz, Joel L., and Wolfgang Härdle. 1996. "Direct Semiparametric Estimation of Single-Index Models with Discrete Covariates." *J. Amer. Statist. Assoc.* 91 (436): 1632–40. https://doi.org/10.2307/2291590.

Hyvärinen, Aapo. 2005. "Estimation of Non-Normalized Statistical Models by Score Matching." *J. Mach. Learn. Res.* 6: 695–709.

References IV

Ichimura, Hidehiko. 1993. "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models." *J. Econometrics* 58 (1-2): 71–120. https://doi.org/10.1016/0304-4076(93)90114-K.

Janzamin, Majid, Hanie Sedghi, and Anima Anandkumar. 2014. "Score Function Features for Discriminative Learning: Matrix and Tensor Framework." *arXiv Preprint arXiv:1412.2863*.

Li, Ker-Chau. 1991. "Sliced Inverse Regression for Dimension Reduction." *J. Amer. Statist. Assoc.* 86 (414): 316–42. http://links.jstor.org/sici?sici=0162-1459(199106)86:414<316:SIRFDR>2.0.CO;2-V&origin=MSN.

Ma, Shujie, and Peter X.-K. Song. 2015. "Varying Index Coefficient Models." *J. Amer. Statist. Assoc.* 110 (509): 341–56. https://doi.org/10.1080/01621459.2014.903185.

Plan, Yaniv, and Roman Vershynin. 2016. "The Generalized Lasso with Non-Linear Observations." *IEEE Trans. Inform. Theory* 62 (3): 1528–37. https://doi.org/10.1109/TIT.2016.2517008.

Sharpe, William F. 1963. "A Simplified Model for Portfolio Analysis." *Management Science* 9 (2). INFORMS: 277–93.

References V

Stein, Charles, Persi Diaconis, Susan Holmes, and Gesine Reinert. 2004. "Use of Exchangeable Pairs in the Analysis of Simulations." In *Stein's Method: Expository Lectures and Applications*, 46:1–26. IMS Lecture Notes Monogr. Ser. Inst. Math. Statist., Beachwood, OH. https://doi.org/10.1214/lnms/1196283797.

Stock, James H, and Mark W Watson. 1988. "A Probability Model of the Coincident Economic Indicators." National Bureau of Economic Research Cambridge, Mass., USA.

Xia, Yingcun, Howell Tong, and W. K. Li. 2002. "Single-Index Volatility Models and Estimation." *Statist. Sinica* 12 (3): 785–99.

Xia, Yingcun, Howell Tong, W. K. Li, and Li-Xing Zhu. 2002. "An Adaptive Estimation of Dimension Reduction Space." *J. R. Stat. Soc. Ser. B Stat. Methodol.* 64 (3): 363–410. https://doi.org/10.1111/1467-9868.03411.

Xue, Liugen, and Qihua Wang. 2012. "Empirical Likelihood for Single-Index Varying-Coefficient Models." *Bernoulli* 18 (3): 836–56. https://doi.org/10.3150/11-BEJ365.

References VI

Yang, Zhuoran, Krishnakumar Balasubramanian, and Han Liu. 2017. "High-Dimensional Non-Gaussian Single Index Models via Thresholded Score Function Estimation." In *Proceedings of the 34th International Conference on Machine Learning*, edited by Doina Precup and Yee Whye Teh, 70:3851–60. Proceedings of Machine Learning Research. International Convention Centre, Sydney, Australia: PMLR. http://proceedings.mlr.press/v70/yang17a.html.

Zhang, Hongfan. 2018. "Quasi-Likelihood Estimation of the Single Index Conditional Variance Model." *Comput. Statist. Data Anal.* 128: 58–72. https://doi.org/10.1016/j.csda.2018.06.008.