

Sliced Regression for Dimension Reduction

HANSHENG WANG¹ AND YINGCUN XIA²

¹GUANGHUA SCHOOL OF MANAGEMENT, PEKING UNIVERSITY

²DEPARTMENT OF APPLIED PROBABILITY AND STATISTICS,

NATIONAL UNIVERSITY OF SINGAPORE

The Inverse Regression Method

- Sliced Inverse Regression (Li, 1991)
- Sliced Average Variance Estimation (Cook and Weisberg, 1991)
- Principal Hessian Direction (Li, 1991; Cook, 1998).
- Inverse Regression Estimation (Cook and Ni, 2005).
- Contour Regression (Li, et al., 2005).
- Fourier Method (Zhu and Zeng, 2006).
- Directional Regression (Li and Wang, 2007).

The Inverse Regression Method

- The merits:
 - computationally highly efficient.
 - might be efficient under a inverse regression setup (i.e., data is generated as $X|Y$).
 - might be insensitive to outliers in response (e.g., SIR and SAVE, but not PHD).
 - might handle both continuous and discrete response in a unified manner by slicing.
- The problems:
 - Depends on linearity condition and/or constant variance assumptions.
 - Highly inefficient under a regression setup (i.e., data is generated as $Y|X$).
 - Might not be exhaustive in the CS estimation.

The Semiparametric Regression Method

- Projection Pursuit Regression (Friedman and Stuetzle, 1981, JASA).
- Minimal Average Variance Estimation (Xia, et al., 2002, JRSSB).
- Partially Linear Multiple Index Model (Samarov, et al., 2005, JASA).
- A Constructive Approach (Xia, 2008, Annals).

The Semiparametric Regression Method

- The merits:
 - Highly efficient under a regression setup (i.e., data is generated as $Y|X$).
 - Might be able to estimate the structural dimension in a data driven manner (e.g., a CV criterion).
 - no distributional assumption is needed for X .
- The problems:
 - Computationally less efficient.
 - Might be sensitive to outliers in the response.
 - Might not be able to estimate the CS exhaustively.
 - Might not be able to handle discrete responses.

An Ideal Estimation Method

- Computationally simple yet efficient.
- Free of linearity condition & constant variance assumption.
- Less sensitive to outliers in predictor.
- Is able to estimate d automatically.
- High estimation efficiency.
- Applicable for both continuous and discrete data.
- Can estimate the CS exhaustively.

The Central Subspace

- Consider random vector (X, Y) with $X \in \mathbb{R}^p$.
- We assume that $Y \perp\!\!\!\perp X | X^\top B$ with $B \in \mathbb{R}^{p \times d}$.
- We call $\mathcal{S}(B)$ an effective dimension reduction (EDR) subspace (Li, 1991), or sufficient dimension reduction (SDR) subspace (Cook, 1998).
- Obviously, $\mathcal{S}(BC)$ is a SDR subspace as long as $C \in \mathbb{R}^{d \times d}$ is of full rank.
- Obviously, $\mathcal{S}(\mathbb{R}^p)$ is a SDR subspace.
- Thus, we are only interested in the smallest SDR subspace, which is defined to be the intersection of all SDR subspaces. We refer to it as the Central Subspace (CS) and denote it by $\mathcal{S}_{Y|X} = \mathcal{S}(B_0)$.

The Central Mean Subspace

- We assume that $E(Y|X) \perp\!\!\!\perp X|X^\top B$ with $B \in \mathbb{R}^{p \times d}$.
- We call $\mathcal{S}(B)$ a mean dimension reduction (MDR) subspace (Cook and Li, 2002).
- Obviously, $\mathcal{S}(BC)$ is a MDR subspace as long as $C \in \mathbb{R}^{d \times d}$ is of full rank.
- Obviously, $\mathcal{S}(\mathbb{R}^p)$ is a MDR subspace.
- Thus, we are only interested in the smallest MDR subspace, which is defined to be the intersection of all MDR subspaces.
- Remark: The intersection of all SDR/MDR subspaces is not guaranteed to be a SDR/MDR subspace. Nevertheless, this is true under rather mild condition.

Two Motivating Propositions

Let $M(y|x) = E\{I(Y \leq y)|X = x\} = G(y|B_0^\top x)$ and $G(y|u) = E(I(Y < y)|B_0^\top X = u)$. Define the gradients $\nabla M(y|x) = (\partial M(y|x)/\partial x_1, \dots, \partial M(y|x)/\partial x_p)^\top$ and $\nabla G(y|u) = (\partial G(y|u)/\partial u_1, \dots, \partial G(y|u)/\partial u_{d_0})^\top$ with $u = (u_1, \dots, u_{d_0})^\top$, we then have

- **Proposition 1.** For any matrix B , $Y \perp\!\!\!\perp X|B^\top X$ is equivalent to $P(Y \leq y|X = x) = P(Y \leq y|B^\top X = B^\top x)$ for all $y \in \mathbb{R}^1$ and $x \in \mathbb{R}^p$.
- **Proposition 2.** Let $\Omega(y) = E\{\nabla M(y|X) \nabla^\top M(y|X)\}$ and $\Lambda(y) = E\{\nabla G(y|B_0^\top X) \nabla^\top G(y|B_0^\top X)\}$. If B_0 is a basis of the CS and that $\nabla M(y|x)$ is continuous in x , then (i) $E\Omega(Y) = B_0 E\{\Lambda(Y)\} B_0^\top$, and (ii) $E\{\Lambda(Y)\}$ is of full rank.

The Slicing Regression Method

- Define a finite number of pre-specified slices, whose grid points are given by

$$\mathcal{T} = \{-\infty = s_0 < s_1 < \cdots < s_H = +\infty\}.$$

- Define the slice indicator as $z_k = I(s_{(k-1)} < Y \leq s_k)$. Theoretically, if the grid points in \mathcal{T} are sufficiently dense, the CMS of $(z_1, \cdots, z_H)^\top \in \mathbb{R}^H$ is expected to coincide with the CS of Y (i.e., $\mathcal{S}_{y|x}$).
- Define a working regression model as

$$z_k = G_k(B_0^\top X) + \epsilon_k, \quad 1 \leq k \leq H,$$

where $G_k(u) = E(z_k | B_0^\top X = u)$ and $\epsilon_k = z_k - G_k(B_0^\top X)$ with $E(\epsilon_k | X) = 0$.

The Slicing Regression Method

- By Proposition 2, $S_{y|x}$ can be estimated consistently and exhaustively through the CMS of z_k .
- The CMS of z_k can be estimated efficiently by many existing methods such as MAVE (Xia et al, 2002) and the methods in Yin and Cook (2002).

Remark – (2.1)

- Note that $G_k(B_0^\top x)$ is related to the conditional distribution function while Xia (2007) considered the conditional density function.
- As a comparison, SR is computationally easier and theoretically more general (e.g., SR is still applicable with discrete responses).
- Furthermore, as we shall demonstrate later, SR enjoys a cross-validation (CV) method, which is able to estimate d_0 consistently.

Remark – (2.2)

- For some inverse regression methods (e.g., SIR and IR), each slice can provide only one directional estimate.
- Consequently, requiring H (i.e., the number of slices) to be greater than d_0 (i.e., the CS dimension) becomes necessary.
- Nevertheless, such a requirement could be problematic if the response is discrete.
- SR is free of such a problem.

Remark – (2.3)

Similar to SIR, the SR approach here considers only the order (or the rank) of the responses rather than their exact values. By doing so, the effect of extreme values or outliers is abated Cavanagh and Sherman (1998). Such a property is another advantage over the traditional MAVE method in terms of the robustness.

An Initial Estimator: OPG

Firstly, minimize the following local least squared function

$$n^{-1} \sum_{i=1}^n \left\{ z_{ik} - a_{jk} - b_{jk}^{\top} X_{ij} \right\}^2 K_{h_0}(X_{ij})$$

Secondly, estimate the OPG matrix by

$$\hat{\Sigma} = n^{-1} \sum_{k=1}^H \sum_{j=1}^n \hat{\rho}_j \hat{b}_{jk} \hat{b}_{jk}^{\top},$$

where $\hat{\rho}_j$ is a trimming function introduced here for technical purpose (Xia, et al., 2002; Fan, et al., 2003).

Remark – (2.3)

- Note that method relies on a working model for the conditional probability function.
- Thus, the simple rule of thumb (or the so-called normal-reference method) can be used to select the bandwidth h_0 (Silverman, 1986; Scott, 1992; Fan and Gijbels, 1996; Li and Racine, 2006).
- Simply speaking, after standardizing the covariate (see Remark 2.6), we set $h_0 = n^{-1/(p+4)}$ throughout the rest of the article.

Remark – (2.4)

Intuitively, those points with too few observations around cannot produce reliable estimates (e.g., \hat{a}_{jk} and \hat{b}_{jk}). Thus, those estimates should be trimmed off. For such a purpose, we define in this article $\hat{\rho}_j = \rho(\hat{f}(X_j))$, where \hat{f} is some estimate of the predictor density and $\rho(\cdot)$ is a function, such that $\rho(\omega) > 0$ if $\omega > \omega_0$, and $\rho(\omega) = 0$ if $\omega \leq \omega_0$ for some small $\omega_0 > 0$. For a more detailed discussion, one can refer to Xia et al. (2002) and Fan, et al. (2003).

The Refined Estimator

Given a current estimate $B_{(t)}$, the next (i.e., refined) estimate $B_{(t+1)}$ can be obtained by minimizing the following global least squares function (Xia, et al., 2002)

$$n^{-2} \sum_{k=1}^H \sum_{j=1}^n \hat{\rho}_j \sum_{i=1}^n \left\{ z_{ik} - a_{jk} - d_{jk}^\top B^\top X_{ij} \right\}^2 K_{h(t)}(X_{ij}^\top B_{(t)})$$

with respect to $a_{jk} \in \mathbb{R}^1$, $d_{jk} \in \mathbb{R}^{d_0}$, and $B \in \mathbb{R}^{p \times d_0}$ with $B^\top B = I_{d_0}$, where I_{d_0} stands for a d_0 -dimensional identity matrix; see Xia, et al. (2002) for detailed algorithm.

Comments – (2.6) and (2.7)

- (2.6) In real application, we typically standardize X_i by setting $X_i := S_x^{-1/2}(X_i - \bar{X})$, where $\bar{X} = n^{-1} \sum_{i=1}^n X_i$ and $S_x = n^{-1} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^\top$. Then the CS directions should be estimated by $S_x^{-1/2} \hat{B}$.
- (2.7) One can also provide a refined OPG estimator by replacing the high dimensional kernel weight with the refined low dimensional one. We refer to such an estimator as the refined OPG estimator (Xia, et al., 2002; Xia, 2006). Our extensive numerical experience suggests that the performance of rOPG can be comparable but not as good as SR, which corroborates the theoretical findings of Xia (2006).

Determining the CS Dimension

- Firstly, we compute the leaving-out-one estimate for $E(z_k|X)$ as

$$a_{jk,d} = \frac{\sum_{i \neq j} K_{\tilde{h}_d}(\hat{B}_d^\top X_{ij}) z_{ik}}{\sum_{i \neq j} K_{\tilde{h}_d}(\hat{B}_d^\top X_{ij})}, \quad k = 1, \dots, H,$$

where $\tilde{h}_d > 0$ is the final bandwidth (i.e., \tilde{h}) used for \hat{B}_d .

- Next, we define the corresponding CV value as

$$CV(d) = n^{-1} \sum_{k=1}^H \sum_{j=1}^n w(X_j) (z_{jk} - a_{jk,d})^2,$$

where $w(x)$ is another trimming function.

Determining the CS Dimension

- To include the trivial case that X is independent of Y (i.e. $d_0 = 0$), we define

$$CV(0) = n^{-1} \sum_{k=1}^H \sum_{j=1}^n w(X_i) (z_{jk} - \bar{z}_{k,-j})^2 \text{ with}$$
$$\bar{z}_{k,-j} = (n-1)^{-1} \sum_{i \neq j} z_{ik}.$$

- Then, d_0 can be estimated by

$$\hat{d} = \operatorname{argmin}_{0 \leq d \leq d_{\max}} CV(d),$$

where d_{\max} is a pre-specified maximum CS dimension (e.g., $d_{\max} = p$ or $d_{\max} = 5$).

Theoretical Results

- Theorem 1. Suppose conditions (C1)-(C5) in the Appendix B hold, $d = d_0$, and the final bandwidth is \hbar , then the SR estimator \hat{B} is consistent with

$$|\hat{B}\hat{B}^\top - B_0B_0^\top| = O_p\{\hbar^4 + \log n/(n\hbar^{d_0}) + n^{-1/2}\}.$$

Thus, \sqrt{n} -consistency can be achieved with $\hbar \propto n^{-1/(d_0+4)}$ and $d_0 \leq 3$. Thus, no under-smoothing is needed for \hbar !

- Theorem 2. Suppose conditions (C1)-(C5) in the Appendix B hold. Moreover, the bandwidth \hbar_d used for different dimension d satisfies $\hbar_d \propto n^{-1/(d+4)}$. Then,

$$P(\hat{d} = d_0) \rightarrow 1 \text{ as } n \rightarrow \infty.$$

Remark 2.8

- To achieve the \sqrt{n} -consistency with $d_0 > 3$, we can apply one dimensional MAVE to z_{ik} , which produces one directional estimate $\hat{\theta}_k \in \mathbb{R}^p$ for every slice $1 \leq k \leq H$. Then, B_0 can be estimated by the first d_0 eigenvectors of $\sum_{k=1}^H \hat{\theta}_k \hat{\theta}_k^\top$.
- One can show that such an estimate is \sqrt{n} -consistent as long as X_i is normally distributed.
- As we discussed before, this approach may not be able to recover the CS exhaustively. Furthermore, its finite sample performance is not attractive since only one direction is retrieved from each slice.

Simulation Study: Tuning Parameters

- (1) The number of slices for SIR, SAVE, and SR is fixed to be 5 or 10;
- (2) The percentage of empirical directions used by SCR is given by 5% or 10% (Li, et al., 2005);
- (3) For the Fourier method, we fix $\sigma_T^2 = 1.0$ but $\sigma_W^2 = 5\%$ or 10% (Zhu and Zeng, 2006);
- (4) Lastly, the Gaussian kernel is used for MAVE and SR. After standardizing the covariates, the final bandwidth is given by $\tilde{h}_d = \kappa \tilde{h}_d^*$, where $\tilde{h}_d^* = 0.1n^{-1/(d+4)}$. Then, κ is fixed to be 5 or 10 for SR and MAVE.
- (5) All the trimming functions are set to be constants.
- (6) $\Delta(\bar{B}, B_0) = |\bar{B}(\bar{B}^\top \bar{B})^{-1} \bar{B}^\top - B_0(B_0^\top B_0)^{-1} B_0^\top|$.

Simulation Study: Predictor Distribution

- $X_i = \Sigma_x^{1/2} e_i$, where Σ_x is a positive definite matrix with its (j_1, j_2) th entry being $0.5^{|j_1 - j_2|}$.
- Moreover, e_i is generated as $e_i = (e_{i1}, \dots, e_{ip})^\top$ with e_{ij} 's independently generated from
 - (N) a standard normal distribution
 - (U) a uniform distribution on $[-\sqrt{3}, +\sqrt{3}]$
 - (M) a mixture normal distribution; generated according to $N(\mu_j, \Sigma_x)$, $j = 1, \dots, p$ with probability $1/p$ each, where $\mu_j \in \mathbb{R}^p$ is a p -dimensional predictor with the j th component being 2 and others 0.
- Number of simulation iteration = 100.

Simulation Study: Examples

Model 1: $Y = (x_1 + x_2 + x_3 + x_4)^{-1} + 0.2 \times \varepsilon$

Model 2: $Y = \cos(2x_1) - \cos(x_2) + 0.2 \times \varepsilon$

Model 3: $Y = \frac{x_1}{0.5 + (x_2 + 1.5)^2} + x_3^2 \times \varepsilon$

Model I: $Y = (x_1 + x_2 + x_3 + x_4)^{-1} + 0.2 \times \varepsilon$

Model II: $Y = 0.1(x_1 + x_2 + x_3 + x_4 + \varepsilon)^3,$

Model III: $Y = \exp(x_1 + 0.5x_2 + x_3) \times \varepsilon,$

Model IV: $Y = \text{sign}(2x_2 + \varepsilon) \times \log |2x_{i2} + 4 + \varepsilon_{i2}|$

Table 1: The Mean (Standard Deviation) of the Estimation Errors with $p = 10$

Predictor Distribution	Tuning Parameter*	SIR	PHD	SAVE	Fourier	SCR	MAVE	SR
Example 1								
N	5	0.31 (0.073)	1.00 (0.004)	0.24 (0.069)	0.78 (0.172)	0.47 (0.158)	0.99 (0.014)	0.08 (0.020)
	10	0.24 (0.068)		0.28 (0.085)	0.80 (0.175)	0.49 (0.161)	0.99 (0.014)	0.06 (0.015)
U	5	0.30 (0.075)	1.00 (0.004)	0.23 (0.069)	0.80 (0.150)	0.46 (0.156)	0.99 (0.014)	0.08 (0.026)
	10	0.24 (0.063)		0.26 (0.086)	0.79 (0.166)	0.48 (0.160)	0.99 (0.059)	0.06 (0.017)
M	5	0.23 (0.065)	0.99 (0.021)	0.19 (0.049)	0.80 (0.202)	0.39 (0.132)	0.99 (0.020)	0.07 (0.018)
	10	0.22 (0.060)		0.24 (0.079)	0.87 (0.171)	0.41 (0.138)	0.99 (0.015)	0.05 (0.014)
Example 2								
N	5	0.97 (0.034)	0.39 (0.089)	0.39 (0.096)	0.79 (0.175)	0.77 (0.166)	0.07 (0.023)	0.19 (0.061)
	10	0.98 (0.041)		0.43 (0.100)	0.58 (0.187)	0.74 (0.170)	0.07 (0.019)	0.18 (0.064)
U	5	0.97 (0.050)	0.50 (0.117)	0.41 (0.107)	0.89 (0.133)	0.98 (0.023)	0.06 (0.015)	0.14 (0.041)
	10	0.98 (0.040)		0.47 (0.120)	0.77 (0.191)	0.98 (0.033)	0.06 (0.016)	0.14 (0.041)
M	5	0.95 (0.068)	0.54 (0.194)	0.55 (0.205)	0.83 (0.155)	0.81 (0.168)	0.06 (0.093)	0.16 (0.043)
	10	0.96 (0.062)		0.68 (0.193)	0.69 (0.192)	0.79 (0.185)	0.06 (0.093)	0.21 (0.218)
Example 3								
N	5	0.93 (0.077)	0.92 (0.092)	0.86 (0.144)	0.68 (0.177)	0.71 (0.208)	0.85 (0.158)	0.33 (0.136)
	10	0.92 (0.098)		0.92 (0.105)	0.68 (0.185)	0.70 (0.206)	0.84 (0.158)	0.34 (0.150)
U	5	0.93 (0.082)	0.91 (0.080)	0.84 (0.142)	0.59 (0.130)	0.68 (0.161)	0.82 (0.153)	0.32 (0.100)
	10	0.94 (0.072)		0.90 (0.114)	0.56 (0.121)	0.69 (0.169)	0.81 (0.167)	0.33 (0.132)
M	5	0.91 (0.098)	0.91 (0.113)	0.89 (0.129)	0.55 (0.127)	0.70 (0.173)	0.84 (0.166)	0.26 (0.072)
	10	0.87 (0.136)		0.92 (0.102)	0.59 (0.153)	0.69 (0.182)	0.87 (0.134)	0.27 (0.085)

*Note: The tuning parameter is: (1) the number of slices for SIR, SAVE, and SR; (2) the σ_W^2 value for Fourier in %; (3) the proportion of empirical directions for SCR in %; (4) the κ value for MAVE; (5) irrelevant for PHD.

Table 2: The Simulation Results of CV

Example	n	Predictor Distribution	Percentage of Estimated Structure Dimension					
			0	1	2	3	4	5
1	100	N	0.00 (0.000)	0.23 (0.423)	0.45 (0.500)	0.24 (0.429)	0.08 (0.273)	0.00 (0.000)
		U	0.00 (0.000)	0.25 (0.435)	0.44 (0.499)	0.20 (0.402)	0.09 (0.288)	0.02 (0.141)
		M	0.00 (0.000)	0.27 (0.446)	0.42 (0.496)	0.26 (0.441)	0.02 (0.141)	0.03 (0.171)
	200	N	0.00 (0.000)	0.84 (0.368)	0.15 (0.359)	0.01 (0.100)	0.00 (0.000)	0.00 (0.000)
		U	0.00 (0.000)	0.75 (0.435)	0.20 (0.402)	0.04 (0.197)	0.01 (0.100)	0.00 (0.000)
		M	0.00 (0.000)	0.79 (0.409)	0.20 (0.402)	0.01 (0.100)	0.00 (0.000)	0.00 (0.000)
	400	N	0.00 (0.000)	1.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)
		U	0.00 (0.000)	1.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)
		M	0.00 (0.000)	1.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)
2	100	N	0.00 (0.000)	0.02 (0.141)	0.19 (0.394)	0.38 (0.488)	0.27 (0.446)	0.14 (0.349)
		U	0.00 (0.000)	0.02 (0.141)	0.26 (0.441)	0.51 (0.502)	0.13 (0.338)	0.08 (0.273)
		M	0.00 (0.000)	0.00 (0.000)	0.30 (0.461)	0.30 (0.461)	0.27 (0.446)	0.13 (0.338)
	200	N	0.00 (0.000)	0.00 (0.000)	0.83 (0.378)	0.11 (0.314)	0.05 (0.219)	0.01 (0.100)
		U	0.00 (0.000)	0.02 (0.141)	0.80 (0.402)	0.15 (0.359)	0.03 (0.171)	0.00 (0.000)
		M	0.00 (0.000)	0.01 (0.100)	0.73 (0.446)	0.18 (0.386)	0.07 (0.256)	0.01 (0.100)
	400	N	0.00 (0.000)	0.00 (0.000)	1.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)
		U	0.00 (0.000)	0.00 (0.000)	1.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)
		M	0.00 (0.000)	0.00 (0.000)	1.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)
3	100	N	0.00 (0.000)	0.00 (0.000)	0.15 (0.359)	0.31 (0.465)	0.36 (0.482)	0.18 (0.386)
		U	0.00 (0.000)	0.02 (0.141)	0.17 (0.378)	0.39 (0.490)	0.24 (0.429)	0.18 (0.386)
		M	0.00 (0.000)	0.00 (0.000)	0.06 (0.239)	0.39 (0.490)	0.42 (0.496)	0.13 (0.338)
	200	N	0.00 (0.000)	0.00 (0.000)	0.14 (0.349)	0.58 (0.496)	0.26 (0.441)	0.02 (0.141)
		U	0.00 (0.000)	0.00 (0.000)	0.08 (0.273)	0.74 (0.441)	0.17 (0.378)	0.01 (0.100)
		M	0.00 (0.000)	0.00 (0.000)	0.05 (0.219)	0.72 (0.451)	0.19 (0.394)	0.04 (0.197)
	400	N	0.00 (0.000)	0.00 (0.000)	0.02 (0.141)	0.94 (0.239)	0.04 (0.197)	0.00 (0.000)
		U	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.96 (0.197)	0.04 (0.197)	0.00 (0.000)
		M	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.94 (0.239)	0.06 (0.239)	0.00 (0.000)

Table 3: The Mean (Standard Deviation) of the Estimation Errors with $p = 10, 20,$ or 50

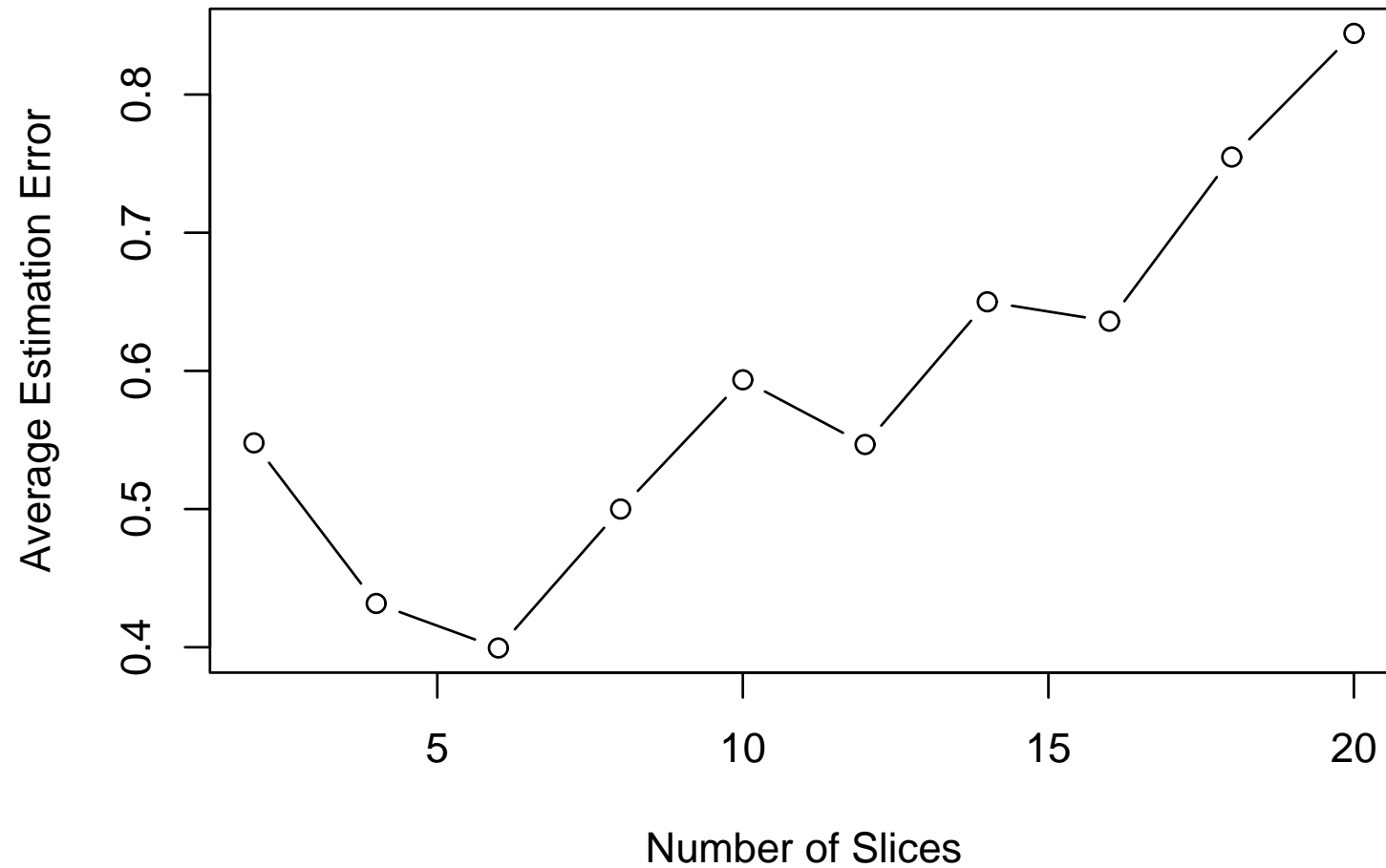
Model	Tuning		SIR	PHD	SAVE	Fourier	SCR	MAVE	SR	
	Parameter*	p								
I	5	10	0.31 (.073)	1.00 (.004)	0.24 (.069)	0.78 (.172)	0.47 (0.158)	0.99 (0.014)	0.08 (0.020)	
		20	0.46 (.063)	1.00 (.003)	0.41 (.080)	0.92 (.101)	0.60 (0.165)	0.99 (0.007)	0.12 (0.023)	
		50	0.64 (.055)	1.00 (.001)	0.76 (.088)	0.98 (.039)	0.79 (0.106)	1.00 (0.003)	0.21 (0.029)	
	10	10	0.24 (.068)		0.28 (.085)	0.80 (.175)	0.49 (0.161)	0.99 (0.014)	0.06 (0.015)	
		20	0.37 (.071)		0.61 (.116)	0.94 (.100)	0.62 (0.165)	1.00 (0.007)	0.09 (0.017)	
		50	0.54 (.062)		0.98 (.022)	0.99 (.015)	0.82 (0.102)	1.00 (0.004)	0.16 (0.025)	
	II	5	10	0.13 (.035)	0.87 (.087)	0.14 (.036)	0.18 (.046)	0.20 (0.047)	0.27 (0.089)	0.13 (0.032)
			20	0.21 (.038)	0.92 (.051)	0.26 (.051)	0.25 (.045)	0.29 (0.050)	0.40 (0.105)	0.19 (0.033)
			50	0.33 (.036)	0.97 (.025)	1.00 (.004)	0.41 (.051)	0.48 (0.055)	0.55 (0.085)	0.31 (0.033)
10		10	0.11 (.028)		0.14 (.042)	0.17 (.045)	0.18 (0.040)	0.23 (0.065)	0.11 (0.033)	
		20	0.17 (.033)		0.67 (.227)	0.25 (.045)	0.26 (0.046)	0.33 (0.056)	0.17 (0.033)	
		50	0.28 (.031)		1.00 (.002)	0.46 (.069)	0.44 (0.051)	0.47 (0.055)	0.27 (0.032)	
III		5	10	0.20 (.054)	0.84 (.096)	0.26 (.076)	0.43 (.115)	0.37 (0.098)	0.81 (0.150)	0.19 (0.048)
			20	0.29 (.057)	0.91 (.050)	0.62 (.177)	0.60 (.119)	0.54 (0.106)	0.92 (0.085)	0.27 (0.059)
			50	0.45 (.051)	0.97 (.022)	1.00 (.005)	0.83 (.079)	0.73 (0.066)	0.98 (0.036)	0.46 (0.057)
	10	10	0.16 (.046)		0.28 (.091)	0.50 (.127)	0.37 (0.090)	0.79 (0.132)	0.14 (0.036)	
		20	0.24 (.047)		0.98 (.037)	0.70 (.123)	0.53 (0.105)	0.89 (0.101)	0.22 (0.041)	
		50	0.39 (.045)		1.00 (.003)	0.92 (.051)	0.73 (0.065)	0.97 (0.036)	0.36 (0.047)	
	IV	5	10	0.27 (.066)	0.68 (.191)	0.65 (.218)	0.24 (.057)	0.26 (0.067)	0.60 (0.190)	0.22 (0.046)
			20	0.37 (.066)	0.87 (.110)	0.96 (.052)	0.34 (.068)	0.38 (0.076)	0.75 (0.151)	0.32 (0.051)
			50	0.56 (.051)	0.99 (.017)	1.00 (.002)	0.52 (.052)	0.58 (0.071)	0.88 (0.092)	0.51 (0.054)
10		10	0.25 (.061)		0.89 (.123)	0.25 (.059)	0.25 (0.064)	0.42 (0.166)	0.21 (0.044)	
		20	0.35 (.070)		0.99 (.017)	0.35 (.069)	0.37 (0.076)	0.61 (0.168)	0.32 (0.057)	
		50	0.54 (.055)		1.00 (.004)	0.56 (.050)	0.56 (0.068)	0.85 (0.118)	0.54 (0.054)	

*Note: The tuning parameter is: (1) the number of slices for SIR, SAVE, and SR; (2) the σ_W^2 value for Fourier in %; (3) the proportion of empirical directions for SCR in %; (4) the κ value for MAVE; (5) irrelevant for PHD.

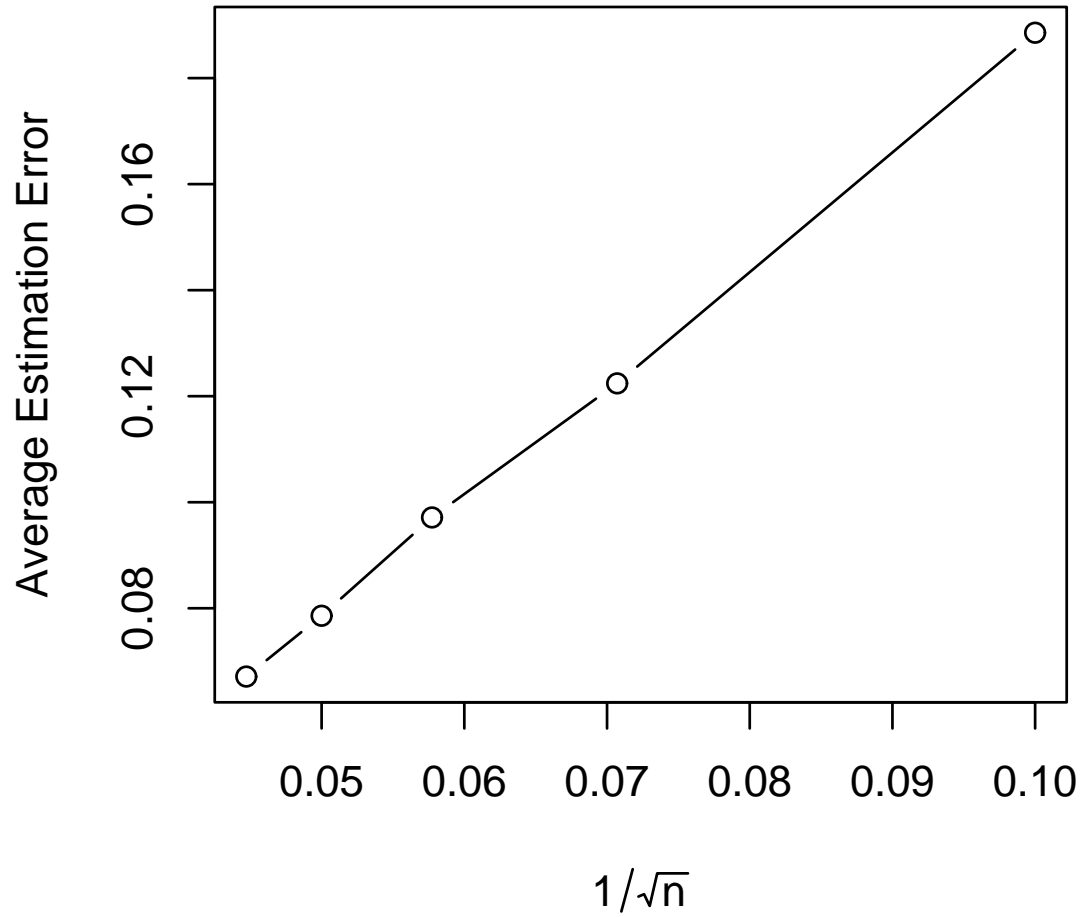
The Effect of Initial Estimate

Model	The Initial Estimate			
	OLS	SIR	SAVE	OPG
I	0.07 (0.018)	0.07 (0.018)	0.07 (0.018)	0.07 (0.018)
II	0.10 (0.021)	0.10 (0.021)	0.10 (0.021)	0.10 (0.021)
III	0.15 (0.035)	0.15 (0.035)	0.15 (0.035)	0.15 (0.035)
IV	0.17 (0.037)	0.17 (0.037)	0.17 (0.037)	0.17 (0.037)

The Effect of H by Model 2



Root- n Consistency by Model 1



The ROE Data: the Dataset

- The cleaned dataset contains a total of 1042 observations collected in the years of 2002 and 2003.
- Each observation corresponds to one firm whose stock is publicly listed on the Chinese stock market during that period.
- For each firm, the following 6 accounting variables are collected in the year of 2002.
 - return on equity (ROE), asset turnover ratio (ATO)
 - profitability margin (PM), leverage level (LEV)
 - sales growth rate (GROWTH), log-transformed total asset (ASSET).
- All predictors are standardized separately so that each of them has unit sample variance.
- The response of interest is the firm's next year earnings (ROEt, i.e., ROE in 2003).

The ROE Data: the Issues

- Because China has experienced a very fast economic growth during the past decade, the firms operating in such an environment also experienced a lot of turbulence and uncertainties. This makes their earnings pattern extremely abnormal and unlikely to be linear.
- For example, the kurtosis estimated based on the residuals differentiated from an ordinary least squares fit is as large as 44.1. Such a heavy-tailed distribution seriously challenges the reliability of those methods that are based on least squares estimation, e.g., MAVE.
- Furthermore, simple histograms reveal that many predictors considered here are highly skewed (e.g., the estimated skewness of ROE is as large as -4.5). Thus, the joint distribution of the predictors cannot be elliptically symmetric.

ROE Data Analysis Results

Variable	SIR	PHD	SAVE	Fourier	SCR	MAVE	SR
ROE	0.528	-0.118	0.984	0.965	0.396	0.103	0.956
ATO	0.283	0.230	-0.012	0.055	0.465	-0.762	0.138
PM	0.792	0.864	-0.074	0.223	0.451	0.505	0.124
LEV	-0.000	-0.211	-0.146	-0.106	-0.284	-0.150	-0.056
GROWTH	0.109	0.345	0.034	0.035	0.354	0.321	0.216
ASSET	0.048	0.149	0.059	0.056	0.211	0.167	0.044

The Estimation Results

- CV estimates $\hat{d} = 1$.
- It clearly detects that a firm's current year earnings (ROE) has the largest positive effect.
- Furthermore, it suggests that the firms with better ATO, PM, and GROWTH tend to have better earnings.
- Lastly, the SR estimate indicates that the effects of LEV and ASSET are very small.
- Both the signs and the magnitude of those estimates match their economics meanings very well.
- As one can see from Figure 1(B), the SR estimate produces a very interesting regression pattern. The part with $\bar{B}^\top X_i > 0$ is approximately linear, but the part with $\bar{B}^\top X_i < 0$ is quite diverse.

Comments on the Best View

- Note that the firms with $\bar{B}^\top X_i < 0$ typically have: negative ROE, poor ATO, little PM, or slow GROWTH.
- Under the pressure of possible severe punishment from the China Security Regulation Commission, those firms may take risks to alternate their normal business operations and even manipulate their earnings report.
- As a consequence, their earnings pattern demonstrates a very high volatility as shown in Figure 1(B).
- In contrast, the firms with $\bar{B}^\top X_i > 0$ suffer much less pressure on such an issue, and tend to maintain their normal business operation. This makes their earnings pattern linear and very predictable; see the top right corner of Figure 1(B).

Other Estimates

- Both the PHD and SAVE estimates produce different signs for ROE and ATO, respectively.
- The problem with the SIR, SCR, and MAVE estimates is that they fail to identify ROE as the most important predictor.
- The only comparable estimation is the Fourier estimate but having very different estimates for GROWTH.
- Question: SR or Fourier, which one is better?

A Rough Comparison

- As one can see from Figures 1(A) and 1(B), both estimates share very similar earnings patterns. Moreover, we do not have a natural measure to compare their goodness of fits.
- Nevertheless, the primary patterns demonstrated by both methods are monotonically increasing, a better estimate is expected to generate higher rank-based correlation coefficient. This motivates us to calculate the the sample correlation coefficients between the ranks of $\bar{B}^\top X_i$ and Y_i .
- We find such a rank-based correlation coefficient is as high as 78.3% for the SR estimate but only 57.2% for the Fourier estimate, implying that the SR estimate might be more accurate.

A Real Data Simulation

- We treat our original sample (with 1042 observations) as if they were the population. For any method, the estimate produced by the whole dataset can be treated as the population parameter.
- We then draw random samples without replacement from the “population” with various sample sizes ($n = 100, 200, \text{ and } 400$). Thereafter, the same type of estimate can be computed based on those random samples.
- The estimation error can then be computed in the same manner as our simulation studies.
- We considered here two different working dimensions, i.e., $d = 1$ and $d = 2$.
- We replicate the experiment for a total of 100 times.

ROE Data Based Simulation Results

d	n	SIR	PHD	SAVE	Fourier	SCR	MAVE	SR
1	100	0.32 (0.162)	0.75 (0.232)	0.46 (0.294)	0.31 (0.195)	0.62 (0.177)	0.94 (0.089)	0.14 (0.060)
	200	0.20 (0.126)	0.72 (0.251)	0.47 (0.307)	0.20 (0.150)	0.45 (0.164)	0.94 (0.104)	0.09 (0.036)
	400	0.13 (0.090)	0.64 (0.263)	0.31 (0.243)	0.12 (0.068)	0.31 (0.121)	0.95 (0.085)	0.06 (0.021)
2	100	0.92 (0.095)	0.65 (0.266)	0.91 (0.117)	0.37 (0.249)	0.73 (0.218)	0.96 (0.045)	0.76 (0.220)
	200	0.87 (0.147)	0.55 (0.277)	0.89 (0.118)	0.20 (0.147)	0.59 (0.253)	0.94 (0.072)	0.75 (0.233)
	400	0.78 (0.193)	0.44 (0.276)	0.82 (0.177)	0.11 (0.046)	0.38 (0.210)	0.92 (0.105)	0.64 (0.230)