Sliced Regression for Dimension Reduction

Hansheng Wang and Yingcun Xia

Peking University & National University of Singapore

Current Version: December 22, 2008.

Abstract

By slicing the region of the response (Li, 1991, SIR) and applying local kernel regression (Xia et al., 2002, MAVE) to each slice, a new dimension reduction method is proposed. Compared with the traditional inverse regression methods, e.g. sliced inverse regression (Li, 1991), the new method is free of the linearity condition (Li, 1991) and enjoys much improved estimation accuracy. Compared with the direct estimation methods (e.g., MAVE), the new method is much more robust against extreme values and can capture the entire central subspace (Cook, 1998b, CS) exhaustively. To determine the CS dimension, a consistent cross-validation (CV) criterion is developed. Extensive numerical studies including one real example confirm our theoretical findings.

KEY WORDS: cross-validation; earnings forecast; minimum average variance estimation; sliced inverse regression; sufficient dimension reduction

†Hansheng Wang is Associate Professor at Department of Business Statistics & Econometrics, Guanghua School of Management, Peking University, Beijing, P. R. China, 100871 (hansheng@gsm.pku.edu.cn). Yingcun Xia is Associate Professor at Department of Statistics and Applied Probability, National University of Singapore, Singapore, 117546 (staxyc@nus.edu.sg).

Acknowledgement. We are very grateful to the Editor, the Associate Editor, and two referees for their careful reading and constructive comments, which leads to a substantial improved manuscript. We are also very grateful to Prof. Bing Li, Prof. Yu Zhu, and Prof. Peng Zeng for sharing with us their computer programs. Wang's research is partially supported by a grant from NSFC (10771006). Xia's research is partially supported by a grant from the Institute of Risk Management and also a grant from National University of Singapore (FRG R-155-000-063-112).

1. INTRODUCTION

Li (1991) developed a seminal sufficient dimension reduction method called sliced inverse regression (SIR). Ever since then, various methods from the inverse regression perspective have been proposed, including sliced average variance estimation (Cook and Weisberg, 1991, SAVE), principal Hessian directions (Li, 1992; Cook, 1998a, PHD), simple contour regression (Li et al., 2005, SCR), inverse regression (Cook and Ni, 2005, IR), the Fourier estimation (Zhu and Zeng, 2006, Fourier), and many others. All those methods were developed under the linearity condition of Li (1991), which assumes that $E(b_1^\top \boldsymbol{X}|b_2^\top \boldsymbol{X})$ is linear in $b_2^\top \boldsymbol{X}$, where $\boldsymbol{X}$ stands for the predictor vector and b_i $(i = 1, 2)$ are two arbitrary vectors. Some of those methods may also need the so-called constant variance assumption (Cook and Weisberg, 1991; Li, 1992; Cook, 1998a; Li et al., 2005). For a good review, we refer to Cook and Ni (2005).

All those inverse regression methods are computationally simple and practically useful. But many of them fail in one way or another to estimate the central subspace (Cook, 1998b, CS) exhaustively. For example, it is well known that PHD (Li, 1992; Cook, 1998a) can only detect nonlinear patterns and it only estimates the directions in the central mean subspace (Cook and Li, 2002; Yin and Cook, 2002, CMS). On the other hand, slicing regression (Duan and Li, 1991), SIR (Li, 1991), and IR (Cook and Ni, 2005) may fail if the regression relationship is highly symmetric (Li, 1992; Cook, 1998b). Furthermore, our experience shows that the finite sample performance of those methods could be poor if the linearity condition of Li (1991) is violated and/or the CS dimension is larger than two. Hence, practical applications call for new methods which have better efficiency and are free of the linearity condition.

As the first attempt, Xia et al. (2002) proposed the minimum average variance estimation (MAVE) method. Unlike the inverse regression methods (e.g., SIR, SAVE, etc), MAVE directly estimates the CMS directions via kernel smoothing techniques. Hence, MAVE is free of the linearity condition. Furthermore, with the refined low-

dimensional kernel weights, MAVE outperforms many other direct estimation methods, where high-dimensional kernel weights are used (Härdle and Stoker, 1989; Samarov, 1993). Consequently, MAVE is useful for both dimension reduction (Xia et al., 2002) and semiparametric modeling (Xia, 2006). However, just like every other statistical method, MAVE has its own limitation. In particular, MAVE is sensitive to extreme values and can infer only about the CMS, which could be very different from the CS (Cook and Li, 2002; Yin and Cook, 2002).

To overcome such a limitation, we propose here a sliced regression (SR) method for dimension reduction. We first slice the response region as Li (1991) did, and then apply the MAVE method (Xia et al., 2002) to each slice. Lastly, by appropriately combining the resulting MAVE estimates from each slice, the new dimension reduction method (Sliced Regression, SR) is created. Compared with the traditional inverse regression methods (e.g., SIR), SR is free of the linearity condition (Li, 1991) and enjoys a much improved estimation accuracy. Compared with the direct estimation methods (e.g., MAVE), SR is much less sensitive to extreme values and can capture the entire CS exhaustively. To determine the CS dimension, a consistent cross-validation (CV) criterion is developed. Extensive numerical studies including one real example confirm our theoretical findings.

The rest of the article is organized as the follows. Section 2 introduces the SR method, including the motivation, the algorithm, and a CV criterion for the CS dimension determination. The asymptotic properties are also investigated in this section. Finite sample performance of SR is evaluated and compared with some existing methods in Section 3 via both simulated and real datasets. A brief discussion about the other aspects of the proposed methods is provided in Section 4.

2. THE SLICED REGRESSION METHOD

2.1. Model and Notations

Let $Y \in \mathbb{R}^1$ be the response of interest and $\boldsymbol{X} = (X_1, \cdots, X_p)^\top \in \mathbb{R}^p$ be the p-dimensional predictor. To simplify the regression relationship, the following model is assumed (Li, 1991; Cook, 1998b)

$$Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{B}^\top \boldsymbol{X}, \tag{2.1}$$

where " $\perp\!\!\!\perp$ " denotes "conditional independence" and $\boldsymbol{B} \in \mathbb{R}^{p \times d}$ is the coefficient matrix. Model (2.1) implies that $\boldsymbol{B}^\top \boldsymbol{X}$ summarizes all the useful information of $\boldsymbol{X}$ about Y. Let $\mathcal{S}(\tilde{\boldsymbol{A}})$ be the linear subspace spanned by the column vectors of an arbitrary matrix $\tilde{\boldsymbol{A}}$. If (2.1) holds, we then refer to $\mathcal{S}(\boldsymbol{B})$ as the sufficient dimension reduction (SDR) subspace (Cook, 1998b). If the intersection of all SDR subspaces is still a SDR subspace, it is called the central subspace (CS), denoted by $\mathcal{S}_{y|x}$ (Cook, 1998b). We assume further that $\mathcal{S}_{y|x}$ exists with a basis being $\boldsymbol{B}_0 \in \mathbb{R}^{p \times d_0}$ for some $0 < d_0 < p$.

As noted by Cook and Li (2002), researchers very often concern only about the conditional mean $E(Y|\boldsymbol{X})$. In that situation, the objective of dimension reduction becomes how to find some basis matrix A such that

$$Y \perp\!\!\!\perp E(Y|\boldsymbol{X}) | \boldsymbol{A}^\top \boldsymbol{X}.$$

We refer to $\mathcal{S}(\boldsymbol{A})$ as the mean dimension reduction (MDR) subspace (Cook and Li, 2002). Similarly, if the intersection of all MDR subspaces is still a MDR subspace, it is referred to as the central mean subspace (Cook and Li, 2002, CMS). As one can see, the CMS is just a subspace of the CS and could be different from the CS.

2.2. The Motivation

Let $\boldsymbol{x} = (x_1, ..., x_p)^\top$ be a non-random vector in $\mathbb{R}^p$ and y be a non-random scalar in $\mathbb{R}^1$. Our SR method is motivated by the following propositions.

Proposition 1. *For any matrix* $\boldsymbol{B}$, $Y \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{B}^\top\boldsymbol{X}$ *is equivalent to* $P(Y \le y|\boldsymbol{X} = \boldsymbol{x}) = P(Y \le y|\boldsymbol{B}^\top\boldsymbol{X} = \boldsymbol{B}^\top\boldsymbol{x})$ *for all* $y \in \mathbb{R}^1$ *and* $\boldsymbol{x} \in \mathbb{R}^p$.

A proof of this proposition can be found in Zeng and Zhu (2007). Since $P(Y \le y|\boldsymbol{X}) = E\{I(Y \le y)|\boldsymbol{X}\}$, Proposition 1 implies that the CS of Y is related closely to the CMS of $I(Y \le y)$. Consequently, as long as the CMS of $I(Y \le y)$ can be estimated for all $y \in \mathbb{R}^1$, one can recover $\mathcal{S}_{y|x}$ easily. Let $M(y|\boldsymbol{x}) = E\{I(Y \le y)|\boldsymbol{X} = \boldsymbol{x}\}$ and $G(y|u) = E(I(Y < y)|\boldsymbol{B}_0^\top\boldsymbol{X} = \boldsymbol{u})$. Because $\boldsymbol{B}_0$ is a basis of the CS, by Proposition 1 we have $M(y|\boldsymbol{x}) = G(y|\boldsymbol{B}_0^\top\boldsymbol{x})$. Consider the gradients $\bigtriangledown M(y|\boldsymbol{x}) = (\partial M(y|\boldsymbol{x})/\partial x_1, ..., \partial M(y|\boldsymbol{x})/\partial x_p)^\top$ and $\bigtriangledown G(y|\boldsymbol{u}) = (\partial G(y|\boldsymbol{u})/\partial u_1, ..., \partial G(y|\boldsymbol{u})/\partial u_{d_0})^\top$ with $\boldsymbol{u} = (u_1, \cdots, u_{d_0})^\top$. We then have

Proposition 2. *Let* $\Omega(y) = E\{\bigtriangledown M(y|\boldsymbol{X})\ \bigtriangledown^\top M(y|\boldsymbol{X})\}$ *and* $\boldsymbol{\Lambda}(y) = E\{\bigtriangledown G(y|\boldsymbol{B}_0^\top\boldsymbol{X})\ \bigtriangledown^\top G(y|\boldsymbol{B}_0^\top\boldsymbol{X})\}$. *If* $\boldsymbol{B}_0$ *is a basis of the CS and that* $\bigtriangledown M(y|\boldsymbol{x})$ *is continuous in* $\boldsymbol{x}$, *then (i)* $E\Omega(Y) = \boldsymbol{B}_0 E\{\boldsymbol{\Lambda}(Y)\}\boldsymbol{B}_0^\top$, *and (ii)* $E\{\boldsymbol{\Lambda}(Y)\}$ *is of full rank.*

A proof of Proposition 2 is given in the Appendix A at the end of this article. Note that $\bigtriangledown M(y|\boldsymbol{x})$ can be estimated easily by nonparametric methods. Thus, $\boldsymbol{B}_0$ can be estimated by the eigenvectors of $\Omega(y)$. Moreover, since matrix $E\{\boldsymbol{\Lambda}(Y)\}$ is of full rank, the CS of Y can be estimated exhaustively via the eigenvectors of $E\{\Omega(Y)\}$. For an easy implementation, we follow the idea of SIR (Li, 1991) and focus on 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 $z_k = I(s_{(k-1)} < Y \le s_k)$. 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., $\mathcal{S}_{y|x}$). To find the CMS for each slice, consider

$$z_k = G_k(\boldsymbol{B}_0^\top\boldsymbol{X}) + \epsilon_k, \quad , k = 1, ..., H, \tag{2.2}$$

where $G_k(u) = E(z_k|\boldsymbol{B}_0^\top\boldsymbol{X} = \boldsymbol{u})$ and $\epsilon_k = z_k - G_k(\boldsymbol{B}_0^\top\boldsymbol{X})$ with $E(\epsilon_k|\boldsymbol{X}) = 0$. By Proposition 2, $\mathcal{S}_{y|x}$ can be estimated consistently and exhaustively through the CMS

of (2.2), which 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 $G_k(\boldsymbol{B}_0^\top \boldsymbol{x})$ 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, if one wishes to recover the CS exhaustively. Nevertheless, such a requirement could be problematic if the response is discrete. For example, if the number of all possible values for the response is even smaller than d_0, those inverse regression methods can no longer estimate the CS exhaustively, while 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 exactly 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; see Example 1 in Section 3.

2.3. An Initial Estimate

Let $\{(\boldsymbol{X}_i, Y_i), i = 1, ..., n\}$ with $\boldsymbol{X}_i = (X_{i1}, \cdots, X_{ip})^\top$ be n random samples from $(\boldsymbol{X}, Y)$. Variables z_{ik} and ϵ_{ik} used in (2.2) can be defined accordingly. Let $K_0(\cdot)$ be a univariate symmetric density function. For any q-dimensional vector $\boldsymbol{v} = (v_1, \cdots, v_q)^\top$, define $K(\boldsymbol{v}) = K(v_1, \cdots, v_q) = K_0(v_1^2 + \cdots + v_q^2)$ and $K_h(\boldsymbol{v}) = h^{-q} K(\boldsymbol{v}/h)$ with a bandwidth $h > 0$. Then, by the method of local linear smoothing, we can estimate

the value of $(G_k(\boldsymbol{x}^\top \boldsymbol{B}_0), \partial G_k(\boldsymbol{x}^\top \boldsymbol{B}_0)/\partial \boldsymbol{x})$ at $\boldsymbol{X}_j$ by $(\hat{a}_{jk}, \hat{\mathbf{b}}_{jk})$, which is obtained by minimizing the following local least squares function (Fan and Gijbels, 1996)

$$n^{-1} \sum_{i=1}^{n} \Big\{ z_{ik} - a_{jk} - \mathbf{b}_{jk}^\top \mathbb{X}_{ij} \Big\}^2 K_{h_0}(\mathbb{X}_{ij}) \tag{2.3}$$

with respect to $a_{jk} \in \mathbb{R}^1$ and $\mathbf{b}_{jk} \in \mathbb{R}^p$, where $\mathbb{X}_{ij} = \boldsymbol{X}_i - \boldsymbol{X}_j$ and $h_0 > 0$ is a bandwidth. The selection of h_0 will be discussed in Remark 2.4. The solution to (2.3) is given by

$$\begin{pmatrix} \hat{a}_{jk} \\ \hat{\boldsymbol{b}}_{jk} \end{pmatrix} = \Big\{ \sum_{i=1}^{n} K_{h_0}(\mathbb{X}_{ij}) \begin{pmatrix} 1 \\ \mathbb{X}_{ij} \end{pmatrix} \begin{pmatrix} 1 \\ \mathbb{X}_{ij} \end{pmatrix}^\top \Big\}^{-1} \Big\{ \sum_{i=1}^{n} K_{h_0}(\mathbb{X}_{ij}) \begin{pmatrix} 1 \\ \mathbb{X}_{ij} \end{pmatrix} z_{ik} \Big\}.$$

Following the idea of the outer-product of gradient method (Samarov, 1993; Xia et al., 2002), we then construct the following matrix

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

where $\hat{\rho}_j$ is a trimming function introduced here for technical purpose (Xia et al., 2002; Fan et al., 2003); see Remark 2.5 for details. Note that $\hat{\boldsymbol{\Sigma}}$ is nothing but an estimator of $E\{\Omega(Y)\}$ in Proposition 2. The basis of $\mathcal{S}_{y|x}$ can then be estimated by the first d_0 eigenvectors of $\hat{\boldsymbol{\Sigma}}$. Such a simple estimate is referred to as outer product of gradient (Xia et al., 2002; Xia, 2006, OPG) estimator, and can be used as one possible initial estimate, denoted by $\boldsymbol{B}_{(0)}$.

Remark 2.4. Note that model (2.2) is just 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 2.5. Intuitively, those points with too few observations around cannot produce reliable estimates (e.g., $\hat{a}_{jk}$ and $\hat{\boldsymbol{b}}_{jk}$). Thus, those estimates should be trimmed off. For such a purpose, we define in this article $\hat{\rho}_j = \rho(\hat{f}(\boldsymbol{X}_j))$, 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).

2.4. The Refined Estimate

The working model (2.2) implies that the gradient vector $\partial G_k(\boldsymbol{x}^\top \boldsymbol{B}_0)/\partial \boldsymbol{x}$ is contained in $\mathcal{S}_{y|\boldsymbol{x}} = \mathcal{S}(\boldsymbol{B}_0)$ for any $\boldsymbol{x} \in \mathbb{R}^p$ and every $1 \leq k \leq H$. Such a relationship suggests that the estimate $\boldsymbol{B}_{(0)}$ can be further refined. Hence, we follow the idea of MAVE (Xia et al., 2002) and propose the following refining procedure. Given a current estimate $\boldsymbol{B}_{(t)}$, the next (i.e., refined) estimate $\boldsymbol{B}_{(t+1)}$ can be obtained by minimizing the following global least squares function

$$n^{-2} \sum_{k=1}^{H} \sum_{j=1}^{n} \hat{\rho}_j \sum_{i=1}^{n} \left\{ z_{ik} - a_{jk} - \boldsymbol{d}_{jk}^\top \boldsymbol{B}^\top \mathbb{X}_{ij} \right\}^2 K_{h_{(t)}}(\mathbb{X}_{ij}^\top \boldsymbol{B}_{(t)}) \tag{2.4}$$

with respect to $a_{jk} \in \mathbb{R}^1$, $\boldsymbol{d}_{jk} \in \mathbb{R}^{d_0}$, and $\boldsymbol{B} \in \mathbb{R}^{p \times d_0}$ with $\boldsymbol{B}^\top \boldsymbol{B} = \boldsymbol{I}_{d_0}$, where $\boldsymbol{I}_{d_0}$ stands for a d_0-dimensional identity matrix. Denote the minimizer of $\boldsymbol{B}$ to (2.4) by $\boldsymbol{B}_{(t+1)}$, which is our refined estimator. Once the estimate $\boldsymbol{B}_{(t+1)}$ converges as t increases, the final SR estimate is obtained, which is denoted by $\hat{\boldsymbol{B}}$.

Note that the minimization problem (2.4) can be solved iteratively with respect to $\{(a_{jk}, \boldsymbol{d}_{jk}), j, k = 1, ..., n\}$ and $\boldsymbol{B}$ separately. As a consequence, we need to solve two quadratic programming problems, each of which has an explicit solution. Define operators $\ell(.)$ and $\mathcal{M}(.)$ respectively as $\ell(\boldsymbol{B}) = (\boldsymbol{\beta}_1^\top, \cdots, \boldsymbol{\beta}_d^\top)^\top$ and $\mathcal{M}(\ell(\boldsymbol{B})) = \boldsymbol{B}$, where $\boldsymbol{\beta}_j \in \mathbb{R}^p$ represents the jth $(1 \leq j \leq d)$ column vector of $\boldsymbol{B}$. For any matrix $\boldsymbol{A}$, we use $|\boldsymbol{A}|$ to denote the maximum singular value of an arbitrary matrix $\boldsymbol{A}$, which is

the Euclidean norm if A is a vector. Then, starting with $t = 1$ and $\boldsymbol{B}_{(1)} = \boldsymbol{B}_{(0)}$, the proposed SR algorithm can be carried out as follows.

Step 1: Let $\boldsymbol{B} = \boldsymbol{B}_{(t)}$ and calculate the solutions of $(a_{jk}, \boldsymbol{d}_{jk})$ to the minimization problem (2.4), which gives

$$\begin{aligned}\begin{pmatrix} a_{jk}^{(t)} \\ d_{jk}^{(t)} \end{pmatrix} &= \left\{ \sum_{i=1}^{n} K_{h_{(t)}}(\boldsymbol{B}_{(t)}^{\top}\mathbb{X}_{ij}) \begin{pmatrix} 1 \\ \boldsymbol{B}_{(t)}^{\top}\mathbb{X}_{ij} \end{pmatrix} \begin{pmatrix} 1 \\ \boldsymbol{B}_{(t)}^{\top}\mathbb{X}_{ij} \end{pmatrix}^{\top} \right\}^{-1} \\ &\quad \times \left\{ \sum_{i=1}^{n} K_{h_{(t)}}(\boldsymbol{B}_{(t)}^{\top}\mathbb{X}_{ij}) \begin{pmatrix} 1 \\ \boldsymbol{B}_{(t)}^{\top}\mathbb{X}_{ij} \end{pmatrix} z_{ik} \right\},\end{aligned}$$

where $h_{(t)} = \max\{\varsigma h_{(t-1)}, \hbar\}$, ς is some constant satisfying $1/2 < \varsigma < 1$, and $\hbar \propto n^{1/(d_0+4)}$ is the final bandwidth, which can be selected by the rule-of-thumb.

Step 2: Let $\rho_j^{(t)} = \rho(\hat{f}_{\boldsymbol{B}_{(t)}}(\boldsymbol{X}_j))$ with $\hat{f}_{\boldsymbol{B}_{(t)}}(\boldsymbol{X}_j) = n^{-1}\sum_{i=1}^{n} K_{h_{(t)}}(\boldsymbol{B}_{(t)}^{\top}\mathbb{X}_{ij})$. Fixing $a_{jk} = a_{jk}^{(t)}$ and $\boldsymbol{d}_{jk} = d_{jk}^{(t)}$, calculate the solution of $\boldsymbol{B}$ or $\ell(\boldsymbol{B})$ to (2.4), which produces

$$\begin{aligned}\boldsymbol{\Gamma}^{(t+1)} &= \Big\{ \sum_{k,j,i}^{n} \rho_j^{(t)} K_{h_{(t)}}(\boldsymbol{B}_{(t)}^{\top}\mathbb{X}_{ij}) \mathbb{X}_{ijk}^{(t)} (\mathbb{X}_{ijk}^{(t)})^{\top} \Big\}^{-1} \\ &\quad \times \sum_{k,j,i}^{n} \rho_j^{(t)} K_{h_{(t)}}(\boldsymbol{B}_{(t)}^{\top}\mathbb{X}_{ij}) \mathbb{X}_{ijk}^{(t)} \{z_{ik} - a_{jk}^{(t)}\},\end{aligned}$$

where $\mathbb{X}_{ijk}^{(t)} = d_{jk}^{(t)} \otimes \mathbb{X}_{ij}$ and "$\otimes$" stands for the Kronecker product.

Step 3: Calculate $\boldsymbol{\Lambda}_{(t+1)} = \{\mathcal{M}(\boldsymbol{\Gamma}^{(t+1)})\}^{\top}\mathcal{M}(\boldsymbol{\Gamma}^{(t+1)})$ and $\boldsymbol{B}_{(t+1)} = \mathcal{M}(\boldsymbol{\Gamma}^{(t+1)})\boldsymbol{\Lambda}_{(t+1)}^{-1/2}$.

Step 4: Check the convergency. If the following discrepancy measure (Li et al., 2005) $|\boldsymbol{B}_{(t+1)}\{\boldsymbol{B}_{(t+1)}^{\top}\boldsymbol{B}_{(t+1)}\}^{-1}\boldsymbol{B}_{(t+1)}^{\top} - \boldsymbol{B}_{(t)}\{\boldsymbol{B}_{(t)}^{\top}\boldsymbol{B}_{(t)}\}^{-1}\boldsymbol{B}_{(t)}^{\top}|$ is smaller than some pre-specified tolerance value (e.g. 10^{-6}), we stop the iteration and output the final SR estimate $\hat{\boldsymbol{B}} := \boldsymbol{B}_{(t+1)}$. Otherwise set $t := t + 1$ and go back to Step 1.

Remark 2.6. In real application, we typically standardize $\boldsymbol{X}_i$ by setting $\boldsymbol{X}_i :=$

$\boldsymbol{S}_x^{-1/2}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})$, where $\bar{\boldsymbol{X}} = n^{-1}\sum_{i=1}^n \boldsymbol{X}_i$ and $\boldsymbol{S}_x = n^{-1}\sum_{i=1}^n (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\top$. Then the CS directions should be estimated by $\boldsymbol{S}_x^{-1/2}\hat{\boldsymbol{B}}$.

Remark 2.7. In fact, we can also provide a refined OPG estimator by replacing the high dimensional kernel weight in (2.3) with the refined low dimensional one. We refer to such an estimator as the refined OPG estimator (Xia et al., 2002; Xia, 2006, rOPG). 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, pp. 1119, Corollary 4.3).

2.5. Estimating the CS Dimension

In practice, one may have little prior knowledge about the true CS dimension d_0. The following cross-validation (CV) criterion can be used in selecting d_0. With a working dimension d and its corresponding SR estimate $\hat{\boldsymbol{B}}_d$, we can calculate the "leave-one-out" fitted value for each observation j $(1 \le j \le n)$ as

$$a_{jk,d} = \frac{\sum_{i\neq j} K_{\hbar_d}(\hat{\boldsymbol{B}}_d^\top \mathbb{X}_{ij}) z_{ik}}{\sum_{i\neq j} K_{\hbar_d}(\hat{\boldsymbol{B}}_d^\top \mathbb{X}_{ij})}, \quad k = 1, ..., H,$$

where $\hbar_d > 0$ is the final bandwidth (i.e., $\hbar$) used for $\hat{\boldsymbol{B}}_d$. Following the idea of Xia et al. (2002), we define the corresponding CV value as

$$CV(d) = n^{-1}\sum_{k=1}^{H}\sum_{j=1}^{n} w(\boldsymbol{X}_j)(z_{jk} - a_{jk,d})^2, \tag{2.5}$$

where $w(x)$ is another trimming function. To include the trivial case that $\boldsymbol{X}$ is independent of Y (i.e. $d_0 = 0$), we define $CV(0) = n^{-1}\sum_{k=1}^{H}\sum_{j=1}^{n} w(\boldsymbol{X}_i)(z_{jk} - \bar{z}_{k,-j})^2$ with $\bar{z}_{k,-j} = (n-1)^{-1}\sum_{i\neq j} z_{ik}$. Then, d_0 can be estimated by

$$\hat{d} = \arg\min_{0\le d\le d_{\max}} CV(d),$$

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

2.6. Theoretical Results

Assume that both $\boldsymbol{B}_0$ and $\hat{\boldsymbol{B}}$ have been standardized such that $\boldsymbol{B}_0^\top \boldsymbol{B}_0 = \boldsymbol{I}_{d_0}$ and $\hat{\boldsymbol{B}}^\top \hat{\boldsymbol{B}} = \boldsymbol{I}_d$. The detailed technical conditions and proof of the following theorems are given in Appendix B, which can be obtained from the authors by request or downloaded from an JASA supplemental material website at:

$$http://www.amstat.org/publications/jasa/supplemental\_materials.$$

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

$$|\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^\top - \boldsymbol{B}_0\boldsymbol{B}_0^\top| = O_p\{\hbar^4 + \log n/(n\hbar^{d_0}) + n^{-1/2}\}.$$

Theorem 1 indicates that the consistency rate of the SR estimator is faster than that of the optimal nonparametric estimator. In particular, if $d_0 \leq 3$, the $\sqrt{n}$-consistency can be achieved by taking $\hbar \propto n^{-1/(d_0+4)}$, which is of the same order as the optimal bandwidth in nonparametric smoothing (Fan and Gijbels, 1996). Consequently, no undersmoothing is needed for SR.

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 $\hat{\boldsymbol{\theta}}_k \in \mathbb{R}^p$ for every slice $1 \leq k \leq H$. Then, $\boldsymbol{B}_0$ can be estimated by the first d_0 eigenvectors of $\sum_{k=1}^H \hat{\boldsymbol{\theta}}_k \hat{\boldsymbol{\theta}}_k^\top$. One can show that such an estimate is $\sqrt{n}$-consistent as long as $\boldsymbol{X}_i$ is normally distributed. As we discussed in Remark 2.2, 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.

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, we have*

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

Theorem 2 confirms theoretically that the proposed CV method is indeed consistent in selecting the CS dimension.

2.7. Some Practical Issues

Per the Associate Editor's kind advice, we would like to discuss here a number of practical issues related to SR's implementation. Firstly, note that SR involves two bandwidths. They are, respectively, h_0 (the bandwidth required by the initial OPG estimator) and $\hbar_d$ (the bandwidth required by the refined estimator). Simultaneously tuning two different bandwidths by the classical method such as CV is computationally expensive. In our calculation, after standardizing X we use the so-called normal-reference method (Silverman, 1986; Scott, 1992; Fan and Gijbels, 1996; Li and Racine, 2006) and set $h_0 = n^{-1/(p+4)}$ and $\hbar = n^{-1/(d+4)}$. It is known that these bandwidths are asymptotically optimal for the estimation of the probability density function when the covariates are jointly standard normal. Although real situation might be different, many researchers still find such a normal-reference method very useful as a quick solution (Silverman, 1986; Scott, 1992; Fan and Gijbels, 1996; Li and Racine, 2006). Our experience with SR further confirms such an observation (see Section 3 for simulation demonstration). Simply speaking, we find that, with such a simple bandwidth selector, SR's performance has already been very competitive. Consequently, if one can estimate SR's optimal bandwidth effectively, SR's finite sample performance might be even better. Nevertheless, how to estimate such an optimal bandwidth effectively is indeed an important yet challenging question for future study.

Another important tuning parameter involved in SR is the number of slices H. One

might expect that a relatively larger H can bring more information from the response into regression. Thus, we can reasonably expect that a relatively larger H can lead to more accurate SR estimates. Nevertheless, a too large H inevitably incurs too many local parameters to be estimated, e.g., a_{jk} and $\boldsymbol{d}_{jk}$ in (2.4). Thus, one can also expect that a too large H is not necessarily a good choice. How to select the optimal H is indeed a challenging question for not only SR but also many inverse regression methods (e.g., SIR and SAVE), and is open for discussion. However, our numerical experience suggests that, within a sensible range (e.g., $H \in [5, 10]$) and with a reasonable sample size, SR is rather robust to H as compared with those inverse regression methods (e.g., SIR and SAVE); see Section 3 for detailed simulation comparison and discussions.

The last important issue is the initial estimate. Because SR only involves low dimensional kernel smoothing, SR does not suffer from the *curse of dimensionality* too much, as long as the initial value can be specified at a reasonable precision. Nevertheless, the OPG estimate developed in Section 2.3 clearly suffers from such a weakness. As noted by the Associate Editor, with a high predictor dimension, the bandwidth required by OPG (i.e., h_0) becomes rather large. This makes the local least squares estimate (2.3) very similar to the ordinary least squares (OLS) estimate, which uses slice indicator z_k as the response and X as the predictor. As kindly raised by the Associate Editor, an interesting question is: what would happen if we just use such a simple OLS estimate as the initial estimate? Example 6 in Section 3 suggests that most likely both the OPG estimate and the simple OLS estimate lead to almost identical final SR estimate. This phenomenon is because the simple OLS estimate very often is also consistent for the CS, as long as the linearity condition is satisfied and the regression relationship is not highly symmetric (Duan and Li, 1991; Li, 1991). If the regression relationship is highly symmetric (e.g., Example 2 in Section 3), the OLS estimate cannot capture any direction in the CS (Duan and Li, 1991), thus cannot serve as a good initial estimate. In this situation, other inverse regression methods

(e.g., SAVE and PHD) are found useful. Moreover, Hall and Li (1993) showed that the linearity condition is still approximately satisfied when the dimension of the X is sufficiently large. Thus, such a simple OLS estimate as suggested by the Associate Editor and many other estimates (e.g., SIR, SAVE) might all be good choices for the initial estimate; see Example 6 in Section 3.

3. NUMERICAL EXPERIMENTS

3.1. Simulation Studies

Simulation studies are conducted to evaluate SR's finite sample performance. For comparison purpose, some existing methods (i.e., SIR, PHD, SAVE, Fourier, SCR, and MAVE) are also evaluated. For every method other than PHD, some tuning parameters are inevitably involved. For example, the number of slices (SIR, SAVE, and SR), the percentage of empirical distribution (SCR), the (σ_T^2, σ_W^2) value (Fourier), and also the bandwidths (MAVE and SR). To evaluate the sensitivity of those methods with respect to the tuning parameter specification, the following simulation configurations are considered: (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 $\hbar_d = \kappa \hbar_d^*$, where $\hbar_d^* = 0.1 n^{-1/(d+4)}$. Then, κ is fixed to be 5 or 10 for SR and MAVE. All the trimming functions are set to be constants. For an arbitrary estimate $\bar{\boldsymbol{B}}$, the estimation accuracy is evaluated by $\Delta(\bar{\boldsymbol{B}}, \boldsymbol{B}_0) = |\bar{\boldsymbol{B}}(\bar{\boldsymbol{B}}^\top \bar{\boldsymbol{B}})^{-1} \bar{\boldsymbol{B}}^\top - \boldsymbol{B}_0(\boldsymbol{B}_0^\top \boldsymbol{B}_0)^{-1} \boldsymbol{B}_0^\top|$ (Li et al., 2005). For each parameter setting, a total of 100 simulation replications are conducted.

Example 1. As our first simulation example, we consider the following non-linear

regression model with additive noise

$$Y_i = (\boldsymbol{X}_i^\top \boldsymbol{B}_0)^{-1} + 0.2 \times \varepsilon_i,$$

where ε_i is a standard normal random variable, $\boldsymbol{X}_i \in \mathbb{R}^{10}$ is a 10-dimensional predictor, and $\boldsymbol{B}_0 = (1, 1, 1, 1, , 0, \cdots, 0)^\top \in \mathbb{R}^{10}$ is the regression coefficient. We know immediately that $d_0 = 1$. Predictors $\boldsymbol{X}_i$ is generated according to $\boldsymbol{X}_i = \boldsymbol{\Sigma}_x^{1/2} \boldsymbol{e}_i$, where $\boldsymbol{\Sigma}_x$ is a positive definite matrix with its (j_1, j_2)th entry being $0.5^{|j_1 - j_2|}$. Moreover, $\boldsymbol{e}_i$ is generated as $\boldsymbol{e}_i = (e_{i1}, \cdots, e_{ip})^\top$ with e_{ij}'s independently generated from either a standard normal distribution (N) or a uniform distribution (U) on $[-\sqrt{3}, +\sqrt{3}]$. Per the Associate Editor's advice, we also considered a mixture normal distribution (M) generated according to $N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_x), j = 1, ..., p$ with probability $1/p$ each, where $\boldsymbol{\mu}_j \in \mathbb{R}^p$ is a p-dimensional predictor with the jth component being 2 and others 0. Various sample sizes (e.g., 100, 200, and 400) are examined. Due to space limitation, we only report here the results with $n = 400$. Because the conditional mean function $E(Y_i|\boldsymbol{X}_i) = (\boldsymbol{X}_i^\top \boldsymbol{B}_0)^{-1}$ tends to produce extreme values around the origin, we expect that MAVE cannot perform well. The simulation results summarized in the top panel of Table 1 indeed confirm such an expectation. In this example, SR stands out as the best method followed by SIR, whose performance, nevertheless, is much worse. Furthermore, SR is much less sensitive towards tuning parameters.

Example 2. It is well known that some dimension reduction methods (e.g., SIR) may fail if the regression relationship is highly symmetric (Cook and Weisberg, 1991). Hence, it is of interest to evaluate SR's performance under such a situation. We borrow the following example from Li (1992),

$$Y_i = \cos(2X_{i1}) - \cos(X_{i2}) + 0.2 \times \varepsilon_i,$$

where $\boldsymbol{X}_i$ and ε_i are generated in the same manner as in Example 1 with $n = 400$. For

this example, we have $d_0 = 2$ and $\boldsymbol{B}_0 = \{(1, 0, 0, \cdots, 0)^\top, (0, 1, 0, \cdots, 0)^\top\} \in \mathbb{R}^{10\times 2}$. The simulation results are summarized in the middle panel of Table 1. We find that SIR completely fails. The performance of PHD and SAVE is reasonable, but much worse than that of MAVE and SR. We find that SR is rather robust towards tuning parameter specification, as compared with other methods (e.g., SIR and SAVE). Lastly, since $\mathcal{S}_{y|x}$ is completely contained in the conditional mean $E(Y|\boldsymbol{X})$, SR's somewhat inferior performance to that of MAVE is not surprising. Although slicing can help us in discovering the complicated CS structure, it may also leads to information loss (Cook and Ni, 2006). Theoretically, we can follow the idea of Cook and Ni (2006) and partially recover those information by replacing the z_k in (2.2) by $z_k Y$. Unless extreme values are involved (e.g., Example 1), SR's estimation accuracy is expected to be further improved.

Example 3. In the previous two examples, the information of $\mathcal{S}_{y|x}$ is completely contained in the conditional mean $E(Y|\boldsymbol{X})$. To take into consideration of the conditional variance $\mathrm{var}(Y|\boldsymbol{X})$, the following model is constructed

$$Y_i = \frac{X_{i1}}{0.5 + (X_{i2} + 1.5)^2} + X_{i3}^2 \times \varepsilon_i,$$

where $\boldsymbol{X}_i$ and ε_i are generated in the same manner as Example 1. For this example, we have $d_0 = 3$, $\boldsymbol{B}_0 = \{(1, 0, 0, \cdots, 0)^\top, (0, 1, 0, \cdots, 0)^\top, (0, 0, 1, \cdots, 0)^\top\} \in \mathbb{R}^{10\times 3}$. With $n = 400$, the results are summarized in the bottom panel of Table 1. For this example, SR stands out as the only method works well with finite sample size. The only competitor Fourier's estimation error is still no less than 0.55. Once again, we find our SR method is rather robust to tuning parameter specification with a relatively large sample size, e.g., $n = 400$.

Example 4. In this example, we would like to evaluate the finite sample performance of our CV method for the CS dimension determination. Data are simulated in the same

manner as the previous three examples. To reduce the computational burden, we fix the maximal dimension $d_{\max} = 5$ with $H = \kappa = 5$. The results are summarized in Table 2. We find that the percentage of experiments with $\hat{d} = d_0$ (i.e., the numbers reported in boldface) quickly approaches 1.00 as the sample size increases. Such a pattern numerically confirms that our CV method is indeed consistent.

Example 5. In this example, we would like to evaluate the finite sample performance of our SR method in the high dimensional situation. Specifically, the data are simulated according to the following four models

$$\begin{aligned}
\text{Model I:} \quad & Y_i = (\boldsymbol{X}_i^\top \boldsymbol{B}_0)^{-1} + 0.2 \times \varepsilon_i, \\
\text{Model II:} \quad & Y_i = 0.1(\boldsymbol{X}_i^\top \boldsymbol{B}_0 + \varepsilon_i)^3, \\
\text{Model III:} \quad & Y_i = \exp(\boldsymbol{X}_i^\top \boldsymbol{B}_0) \times \varepsilon_i, \\
\text{Model IV:} \quad & Y_i = \text{sign}(2X_{i2} + \varepsilon_{i1}) \times \log|2X_{i2} + 4 + \varepsilon_{i2}|,
\end{aligned}$$

where Model I is the same as our Example 1, and Models II – IV are borrowed from Duan and Li (1991), Ni et al. (2005) and Chen and Li (1998), respectively. More specifically, ε_i, ε_{i1}, and ε_{i2} are independent standard normal noise. The predictor $\boldsymbol{X}_i$ is generated according to the normal distribution as specified by Example 1. The sample size is fixed to be $n = 400$, but the number of predictors are given by $p = 10$, 20, or 50 respectively. For both Models I and II, we have $\boldsymbol{B}_0 = (1, 1, 1, 1, 0, 0, \cdots, 0)^\top \in \mathbb{R}^p$. For Model 3, we fix $\boldsymbol{B}_0 = (1, 0.5, 1, 0, 0, \cdots, 0)^\top \in \mathbb{R}^p$. As one can see, $d_0 = 1$ for Models I – III, but $d_0 = 2$ for Model IV with $\boldsymbol{B}_0 = \{(1, 0, 0, \cdots, 0)^\top, (0, 1, 0, \cdots, 0)^\top\} \in \mathbb{R}^{p \times 2}$. As one can see from Table 3, SR demonstrates a very satisfactory finite sample performance even with $p = 50$. For Models II – IV, the performance of SIR is also very competitive, but slightly worse than that of SR.

Example 6. To evaluate SR's sensitivity towards different specifications of the initial estimates, we replicate Example 5 but with the following four different initial estimates:

the ordinary least squares estimate with z_k as the response (OLS, as suggested by the Associate Editor), the SIR estimate, the SAVE estimate, and also the OPG estimate. For this example, we fix $p = 10$, $n = 400$, $\kappa = 5$, and $H = 5$ with mixture predictor distribution (M) given in Example 1. The average estimation errors are reported in Table 4. As one can see, regardless of which initial estimate to use (e.g., OLS, SIR, SAVE, or OPG), the final results are almost identical.

Example 7. To further quantify the effect of the H on SR's estimation accuracy, we replicate Example 2 with $n = 200$, $p = 10$, $\kappa = 10$, and normally distributed predictors. We then compute the SR estimate with various H values ($H = 2, 4, \cdots, 20$), and then evaluate its estimation error in the same manner as Example 2. The calculation results are reported in Figure 1. As one can see, a reasonably larger H value (e.g., $H = 2$ vs. $H = 6$) can leads to better SR estimates with smaller average estimation error. However, an unduly large H can also affect the estimation accuracy adversely (e.g., $H = 6$ vs. $H = 20$). Such a pattern well matches our expectation as discussed in the second paragraph of Section 2.7.

Example 8. One might wonder why the finite sample performance of the SR estimate improves much faster than other methods (e.g., SIR) as n increases. Firstly, we want to kindly remark that the MAVE technique utilized by SR very often generates highly efficient estimate. For example, in a single-index model setup, it has been very well understood (Xia, 2006) that the resulting estimate is the mostly efficient in a semi-parametric sense (Bickel et al., 1993). If the MAVE technique can leads to the mostly efficient estimates for the single-index model, it is not surprising then to find it also performs very well under a multiple-index setup and the general dimension reduction. This might partially explain why SR performs so well. Secondly, the outstanding performance of SR might also partially due to its $\sqrt{n}$-consistency. We reconsider Example 1 with $p = 10$, $\kappa = 5$, $H = 5$, and normally distributed predictors. The sample size n is taken as $n = 100, 200, \cdots, 500$. The average estimation error of the SR estimate is

computed and reported in Figure 2. If the SR estimate is indeed $\sqrt{n}$ consistent as we stated in Theorem 1, we should expect an approximately linear relationship between the averaged estimation error $\Delta(\bar{\boldsymbol{B}}, \boldsymbol{B}_0)$ and $1/\sqrt{n}$. As one can see, Figure 2 clearly confirms such an expectation. Consequently, we know that the speed at which the SR's estimation error improves is not surprising. The same experiments are also conducted for both Example 2 and Example 3 with similar findings.

3.2. The ROE Data

We study here a real example from one of the world's fastest growing capital markets – the Chinese stock market. The dataset is derived from the CCER database, which is considered as one of the most authoritative commercial databases for Chinese stock market (*http://www.ccerdata.com/*). The final 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), and log-transformed total asset (ASSET). These variables serve as our predictors. 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 ultimate goal of this study is to understand those firms' earnings patterns, which can be very useful for investment decision.

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 (Cook and Li, 2002) 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 schewed (e.g., the estimated skewness of ROE is as large as -4.5). Thus, the joint distribution of the predictors cannot be elliptically symmetric. Such an observation rules out the possibility that the linearity condition of Li (1991) can be well satisfied here; see also Eaton (1986) and Cook and Nachtsheim (1994). Hence, the applicability of those inverse regression methods is also very questionable here.

As one can see, such a complicated dataset naturally calls for a sufficient dimension reduction method, which must be free of the linearity condition yet insensitive to possible outliers. As a result, the proposed SR method is a good choice to analyze this dataset. We first apply our CV criterion (2.5) to determine the structure dimension, which estimates $\hat{d} = 1$. We are hesitate to determine the structure dimension using the chi-squares tests offered by those inverse regression methods (e.g., SIR, PHD), since their statistical validity is very questionable for this dataset due to serious linearity condition violation. With $d_0 = 1$, we compute the SR estimate and report it in Table 5. For ease of comparison, we also report the estimates of SIR, PHD, SAVE, Fourier, SCR, and MAVE in the same table. The estimates are adjusted so that the sample correlation coefficient between $\bar{B}^\top \boldsymbol{X}_i$ and Y_i is positive, where $\bar{B}$ stands for an arbitrary estimate (e.g., SIR estimate).

From Table 5, we note first that the SR estimate is very reasonable. It clearly detects that a firm's current year earnings (ROE) has the largest positive effect on its future earnings. Furthermore, it suggests that the firms with better asset turnover ratio (ATO, i.e., better capability to make use of its asset), profit margin (PM, i.e., better capability to realize profit), and growth rate (GROWTH, i.e., a fast growing yet promising company) tend to have better earnings next year. Lastly, the SR estimate indicates that the effects of debt level (LEV) and firm size (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 $\bar{B}^\top \boldsymbol{X}_i > 0$ is approximately linear, but the part with $\bar{B}^\top \boldsymbol{X}_i < 0$ is quite diverse.

We remark that the pattern revealed by the SR estimate can be well explained. Note that the firms with $\bar{B}^\top \boldsymbol{X}_i < 0$ typically have: negative current year earnings (ROE), poor capability to manage its asset (ATO), little profit margin (PM), or slow growth rate (GROWTH). Under the pressure of possible severe punishment from the China Security Regulation Commission (the government body overseeing the stock market; see Wang, 2007), 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 $\bar{B}^\top \boldsymbol{X}_i > 0$ 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 3(B).

Compared with the SR estimate, we find that both the PHD and SAVE estimates produce different signs for ROE and ATO, respectively. This implies that the firms making better use of their asset tend to have worse earnings capability. Such a conclusion clearly contradicts common economic theory. 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, which shares the same signs as the SR estimate for each variable but has very different value for GROWTH. As one can see from Figures 1(A) and 1(B), both estimates (i.e., Fourier and SR) share very similar earnings patterns. The regression relationships demonstrated by both estimates are complicated and contaminated with a lot of extreme values and possible outliers. Hence, we do not have a natural measure to compare their goodness of fits. However, because the primary patterns demonstrated by both methods are monotonically increasing, a better estimate is expected to generate higher rank-based correlation co-

efficient. This motivates us to calculate the the sample correlation coefficients between the ranks of $\bar{B}^\top \boldsymbol{X}_i$ 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.

To further confirm SR's outstanding performance, we follow the Editor's advice and conduct the following bootstrap-type experiment. 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, 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. One might wonder whether $d_0 = 1$ (an estimate by SR) is fair to other inverse regression methods, we considered here two different working dimensions, i.e., $d = 1$ and $d = 2$. We replicate the experiment for a total of 100 times for each parameter setting and summarize the averaged estimation error in Table 6.

Firstly, if the true structure dimension is indeed $d_0 = 1$, both Ye and Weiss (2003) and Zhu and Zeng (2006) demonstrated that the estimation variability associated with the second spurious direction should be large, which in turn makes the overall estimation variability with $d = 2$ large. Thus, we would expect that SR's estimation error with $d = 1$ to be relatively small but $d = 2$ to be relatively large (Ye and Weiss, 2003; Zhu and Zeng, 2006). With $d = 1$, SR's estimation error is rather small and steadily decreases as sample size increases. But, SR's estimation error increases substantially with $d = 2$; see Table 6. Such a pattern further supports our CV estimation result $d_0 = 1$. We find that SIR and SAVE estimates follow similar patterns. From that perspective, SIR and SAVE also support $d_0 = 1$. Similar pattern also holds for SCR. For the Fourier method, if we strictly follows the bootstrap procedure as suggested by Zhu

and Zeng (2006), we find the Fourier method also agree with $d_0 = 1$. If we naïvely apply the chi-square tests (developed under the linearity condition) to the PHD method (Li, 1991; Cook, 1998a), we find the true structure dimension is 0, which is certainly not correct. The problem with the MAVE estimate is that its performance is too poor to make any meaningful conclusion, regardless of whether $d = 1$ or $d = 2$. Overall speaking, it seems that $d_0 = 1$ is the most plausible dimension for this particular dataset. It is also remarkable that, with $d_0 = 1$, the estimation error of SR is substantially smaller than those of all its competitors, which corroborates our simulation findings very well.

To conclude this study, we find that our SR method performs best among its competitors. It outperforms the PHD and SAVE methods by keeping the coefficient signs consistent with their economic meaning. It outperforms the SIR, SCR, and MAVE methods by keeping the estimate interpretable. It also outperforms the Fourier method by delivering more accurate estimates.

Remark 3.1. We would like to kindly remark that this ROE dataset is a case where the linearity condition is violated seriously. Consequently, this is a situation unfavorable to those inverse regression methods. For such a reason, one should not mistakenly take this example as an evidence that inverse regression method always perform poorly in real data analysis. In fact, as demonstrated by many researchers, those inverse regression methods (e.g., SIR and SAVE) perform fairly well as long as the linearity condition is reasonably satisfied (Cook and Yin, 2001). In that situation, we always find SR's performance is comparable or even better. The only purpose of this example, per Editor's important advice, is to demonstrate that SR might still work fairly well in the situation where all other methods fail.

4. CONCLUDING REMARKS

We propose in this article a SR method for sufficient dimension reduction. The new method is carried out by slicing the region of the response (Li, 1991) and then

applying the MAVE method to each slice (Xia et al., 2002). The slicing procedure helps us discovering the complicated CS structure, while MAVE frees us from the linearity condition and also improves the estimation accuracy. We show both theoretically and numerically that SR is able to recover the CS exhaustively, which is another advantage over many existing methods.

Another interesting issue is the $\sqrt{n}$-consistency. Since each slice produces only one directional estimate, SIR achieves the $\sqrt{n}$-consistency. Such an operation is equivalent to assuming a single-index working regression model for (2.2). Similar idea can be applied to SR too, but under similar design assumptions; see the discussion in Remark 2.8. It is remarkable that the $\sqrt{n}$-consistency of SIR is just a theoretical result, and its implication in finite sample is quite limited. For example, our numerical experience suggests that SIR frequently misses important CS directions when $d_0 > 2$, which in turn makes its estimation accuracy very poor; see Example 3.

Finally, we need to point out that we do not claim SR as the only best dimension reduction method. In fact, different favorable circumstances do exist for different dimension reduction methods. However, the overall comparison made in the paper, in terms of both theoretical analysis and numerical studies, suggests that SR is indeed a very good method for dimension reduction.

APPENDIX A. PROOF OF PROPOSITION 2

The conclusion (i) is obvious, hence, its proof is omitted. We only focus on the proof of (ii). If $E\{\boldsymbol{\Lambda}(Y)\}$ is not of a full rank, there must exist a vector $\boldsymbol{\eta}_1 \in \mathbb{R}^{d_0}$ such that $\|\boldsymbol{\eta}_1\| = 1$ but $\boldsymbol{\eta}_1^\top E\{\boldsymbol{\Lambda}(Y)\}\boldsymbol{\eta}_1 = E[\{\boldsymbol{\eta}_1^\top \bigtriangledown G(Y|\boldsymbol{B}_0^\top \boldsymbol{X})\}^2] = 0$, which immediately suggests that

$$\boldsymbol{\eta}_1^\top \bigtriangledown G(Y|\boldsymbol{B}_0^\top \boldsymbol{X}) = 0 \quad a.s. \tag{A.1}$$

Expand $\boldsymbol{\eta}_1$ to $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \cdots, \boldsymbol{\eta}_{d_0}) \in \mathbb{R}^{d_0 \times d_0}$ such that $\boldsymbol{\eta}^\top \boldsymbol{\eta} = \boldsymbol{I}_{d_0}$, where $\boldsymbol{I}_{d_0}$ stands for

a d_0-dimensional identity matrix. Define function $\tilde{G}(y|\boldsymbol{u}) = G(y|\boldsymbol{\eta u})$. It follows then

$$G(Y|\boldsymbol{B}_0^\top \boldsymbol{X}) = G\Big(Y\Big|\boldsymbol{\eta}\{\boldsymbol{B}_0\boldsymbol{\eta}\}^\top \boldsymbol{X}\Big) = G\Big(Y\Big|\boldsymbol{\eta}\tilde{\boldsymbol{B}}_0^\top \boldsymbol{X}\Big) = \tilde{G}\Big(Y\Big|\tilde{\boldsymbol{B}}_0^\top \boldsymbol{X}\Big),$$

where $\tilde{\boldsymbol{B}}_0 = \boldsymbol{B}_0\boldsymbol{\eta}$ is another basis of the CS, i.e., $\mathcal{S}(\tilde{\boldsymbol{B}}_0) = \mathcal{S}(\boldsymbol{B}_0)$. Hence, $\tilde{G}(Y|\tilde{\boldsymbol{B}}_0^\top \boldsymbol{X})$ is nothing but another representation of $G(Y|\boldsymbol{B}_0^\top \boldsymbol{X})$ in terms of $\tilde{\boldsymbol{B}}_0^\top \boldsymbol{X}$. We then study how $\tilde{G}(Y|\cdot)$ varies along the direction $\tilde{\boldsymbol{\beta}}_1^\top \boldsymbol{X} = \boldsymbol{\eta}_1^\top \boldsymbol{B}_0^\top \boldsymbol{X}$. We have

$$\frac{\partial \tilde{G}(Y|\tilde{\boldsymbol{B}}_0^\top \boldsymbol{X})}{\partial(\tilde{\boldsymbol{\beta}}_1^\top \boldsymbol{X})} = \frac{\partial G(Y|\boldsymbol{\eta}\tilde{\boldsymbol{B}}_0^\top \boldsymbol{X})}{\partial(\tilde{\boldsymbol{\beta}}_1^\top \boldsymbol{X})} = \bigtriangledown G(Y|\boldsymbol{B}_0^\top \boldsymbol{X})\boldsymbol{\eta}_1 = 0 \quad a.s.,$$

where the last equality is due to (A.1). This suggests that the function $G(Y|\boldsymbol{B}_0^\top \boldsymbol{X}) = \tilde{G}(Y|\tilde{\boldsymbol{B}}_0^\top \boldsymbol{X})$ never varies along the direction $\tilde{\boldsymbol{\beta}}_1^\top \boldsymbol{X}$. Thus, $G(Y|\boldsymbol{B}_0^\top \boldsymbol{X}) = \tilde{G}(Y|\tilde{\boldsymbol{B}}_0^\top \boldsymbol{X})$, as a function of $\tilde{\boldsymbol{B}}_0\boldsymbol{X}$, only depends on the directions $\tilde{\boldsymbol{\beta}}_j^\top \boldsymbol{X} = \boldsymbol{\eta}_j^\top \boldsymbol{B}_0^\top \boldsymbol{X}$ for $j \geq 2$. This implies that the CMS dimension of $I(Y \leq y)$ is less than d_0. By Proposition 1, such a conclusion further suggests that the CS dimension of $Y|\boldsymbol{X}$ is less than d_0. Such a conclusion contradicts the definition of d_0. Consequently, $E\{\boldsymbol{\Lambda}(Y)\}$ must be of full rank. This completes the proof.

REFERENCES

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993), "Efficient and adaptive estimation in semiparametric models", *Johns Hopkins Series in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD.*

Cavanagh, C. and Shreman, R. P. (1998), "Rank estimators for monotonic index models", *Journal of Econometrics*, 84, 351–381.

Chen, C. H. and Li, K. C. (1998), "Can SIR be as popular as multiple linear regression?" *Statistica Sinica*, 8, 289–316.

Cook, R. D. (1998a), "Principal Hessian directions revisited," *Journal of the American Statistical Association*, 93, 84–94.

— (1998b), *Regression Graphics*, John Wiley, New York, NY.

Cook, R. D. and Li, B. (2002), "Dimension reduction for conditional mean in regression," *The Annals of Statistics*, 32, 455–474.

Cook, R. D. and Ni, L. (2005), "Sufficient dimension reduction via inverse regression: a minimum discrepancy approach," *Journal of the American Statistical Association*, 100, 410–428.

Cook, R. D. and Yin, X. (2001), "Dimension reduction and visualization in discriminant analysis (with discussion)," *Australian & New Zealand Journal of Statistics*, 43, 147–199.

Cook, R. D. and Nachtsheim (1994), "Reweighting to achieve elliptically contoured covariates in regression", *Journal of the American Statistical Association*, 89, 592–599.

Cook, R. D. and Ni, L. (2006), "Using intraslice covariances for improved estimation of the central subspace in regression," *Biometrika*, 93, 65–74.

Cook, R. D. and Weisberg, S. (1991), "Discussion of "Sliced inverse regressioin for dimension reduction"," *Journal of the American Statistical Association*, 86, 28–33.

Duan, N. and Li, K. C. (1991), "Slicing regression: a link-free regression method," *The Annals of Statistics*, 19, 505–530.

Eaton, M. L. (1986), "A characterization of spherical distributions", *Journal of Multivariate Analysis*, 20, 272–276.

Fan, J. and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, Chapman and Hall, New York, NY.

Fan, J., Yao, Q., and Cai, Z. (2003), "Adaptive varying-coefficient models," *Journal of the Royal Statistical Society, Series B*, 65, 57–80.

Hall, P. and Li, K. C. (1993), "On almost linearity of low-dimensional projections from high-dimensional data," *The Annals of Statistics*, 21, 867–889.

Härdle, W. and Stoker, T. M. (1989), "Investigating smooth multiple regression by method of average derivative," *Journal of the American Statistical Association*, 84, 986–995.

Li, B., Zha, H., and Chiaromonte, F. (2005), "Contour regression: a general approach to dimension reduction," *The Annals of Statistics*, 33, 1580–1616.

Li, K.-C. (1991), "Sliced inverse regression for dimension reduction," *Journal of the American Statistical Association*, 86, 316–327.

— (1992), "On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma," *Journal of the American Statistical Association*, 87, 1025–1039.

Li, Q. and Racine, J. S. (2006), *Nonparametric Econometrics*, Princeton: Princeton University Press.

Ni, L, Cook, R. D. and Tsai, C. L. (2005), "A note on shrinkage sliced inverse regression", *Biometrika*, 91, 242–247.

Samarov, A. M. (1993) Exploring regression structure using nonparametric functional estimation. *Journal of the American Statistical Association.* **88**, 836-847.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice and visualization.* John Wiely & Sons, New York.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.

Wang, H. (2007), "A note on iterative marginal optimization: a simple algorithm for maximum rank correlation estimation," *Computational Statistics & Data Analysis*, 51, 2803–2812.

Xia, Y. (2006), "Asymptotic distributions of two estimators of the single-index model," *Econometric Theory*, 22, 1112–1137.

— (2007), "A constructive approach to the estimation of dimension reduction directions," *The Annals of Statistics*, To appear.

Xia, Y., Tong, H., Li, W. K., and Zhu, L. (2002), "An adaptive estimation of dimension reduction space," *Journal of Royal Statistical Society, Series B.*, 64, 363–410.

Ye, Z. and Weiss, R. E. (2003), "Using the bootstrap to selection one of a new class of dimension reduction methods", *Journal of the American Statistical Association*, 98, 968–979.

Yin, X. and Cook, R. D. (2002) Dimension Reduction for the Conditional k-th Moment in Regression. *Journal of Royal Statistical Society, Series B.*, 64, 159–175.

Zeng, P. and Zhu, Y. (2007), "An integral transformation method for estimating the central mean and central subspaces", *Manuscript*.

Zhu, Y. and Zeng, P. (2006), "Fourier Methods for Estimating the Central Subspace and the Central Mean Subspace in Regression," *Journal of the American Statistical Association*, 101, 1638–1651.

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

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)

Table 2: The Simulation Results of CV. The prediction distribution is: (N) normal distribution, (U) uniform distribution, and (M) mixture distribution. The numbers reported in boldface are the percentages of the experiments with $\hat{d} = d_0$.

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. 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.

Model	Tuning Parameter*	p	SIR	PHD	SAVE	Fourier	SCR	MAVE	SR
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)

Table 4: The Mean (Standard Deviation) of the Estimation Errors with Different Initial Estimates

	The Initial Estimate			
Model	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)

Table 5: The 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

Table 6: The Mean (Standard Deviation) of the Estimation Errors Based on the Bootstrap ROE Data

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

Figure 1: The Effect of Slice Number on Estimation Accuracy

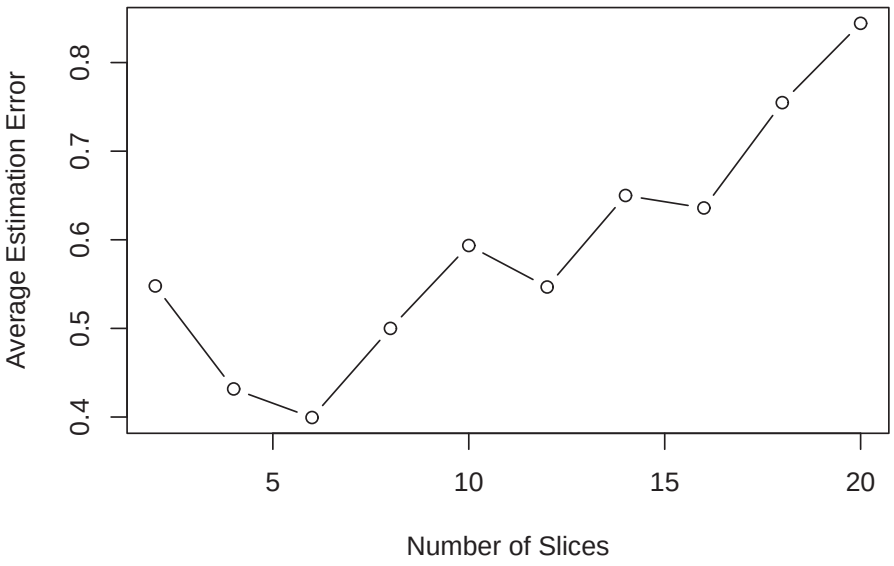

Figure 2: The $\sqrt{n}$-Consistency of the SR Estimate

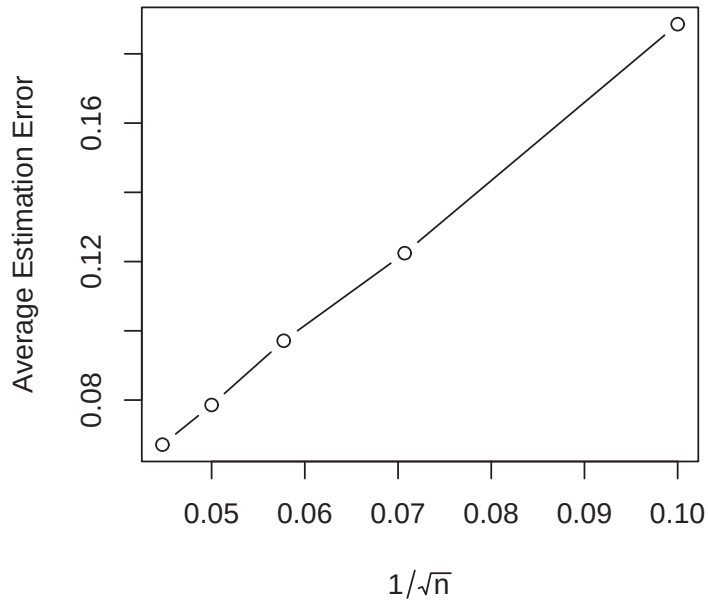

Figure 3: The Scatter Plots of the ROE Data