Journal of Business & Economic Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ubes20

High-Dimensional Elliptical Sliced Inverse Regression in Non-Gaussian Distributions

Xin Chen, Jia Zhang & Wang Zhou

JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2022, VOL. 40, NO. 3, 1204–1215
https://doi.org/10.1080/07350015.2021.1910041

High-Dimensional Elliptical Sliced Inverse Regression in Non-Gaussian Distributions

Xin Chen[a], Jia Zhang[b], and Wang Zhou[c]

[a]Department of Statistics & Data Science, Southern University of Science and Technology, Shenzhen, China; [b]School of Statistics, Southwestern University of Finance and Economics, Chengdu, China; [c]Department of Statistics and Applied Probability, National University of Singapore, Singapore

ABSTRACT

Sliced inverse regression (SIR) is the most widely used sufficient dimension reduction method due to its simplicity, generality and computational efficiency. However, when the distribution of covariates deviates from multivariate normal distribution, the estimation efficiency of SIR gets rather low, and the SIR estimator may be inconsistent and misleading, especially in the high-dimensional setting. In this article, we propose a robust alternative to SIR—called elliptical sliced inverse regression (ESIR), to analysis high-dimensional, elliptically distributed data. There are wide applications of elliptically distributed data, especially in finance and economics where the distribution of the data is often heavy-tailed. To tackle the heavy-tailed elliptically distributed covariates, we novelly use the multivariate Kendall's tau matrix in a framework of generalized eigenvalue problem in sufficient dimension reduction. Methodologically, we present a practical algorithm for our method. Theoretically, we investigate the asymptotic behavior of the ESIR estimator under the high-dimensional setting. Extensive simulation results show ESIR significantly improves the estimation efficiency in heavy-tailed scenarios, compared with other robust SIR methods. Analysis of the Istanbul stock exchange dataset also demonstrates the effectiveness of our proposed method. Moreover, ESIR can be easily extended to other sufficient dimension reduction methods and applied to nonelliptical heavy-tailed distributions.

ARTICLE HISTORY
Received September 2019
Accepted February 2021

KEYWORDS
Central subspace; Convergence rate; Elliptical distribution; Multivariate Kendall's tau; Sliced inverse regression; Sufficient dimension reduction

1. Introduction

In the regression model, let $Y \in \mathbb{R}$ denote the response variable and $\mathbf{X} \in \mathbb{R}^p$ denote the covariates. If there exist orthogonal $p \times 1$ vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ with unit norm such that

$$Y \perp\!\!\!\perp \mathbf{X} \mid (\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}, \ldots, \boldsymbol{\beta}_K^{\mathrm{T}}\mathbf{X}), \quad (K \leq p),$$

where $\perp\!\!\!\perp$ denotes independence, the column space of the $p \times K$ matrix $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$ is defined as a dimension reduction subspace by Cook (1994, 1998). Under mild conditions, the intersection of all the dimension reduction subspaces is still a dimension reduction subspace and is called the central subspace (Cook 1994, 1996). Various methods have been proposed to estimate the central subspace in the literature, which are together referred to as sufficient dimension reduction methods.

Among them, sliced inverse regression (SIR) (Li 1991) is the earliest and most popular method owning to its simplicity, generality and computational efficiency. Li (1991) proved the consistency of SIR for the fixed p setting. Hsing and Carroll (1992) considered the case where each slice only contained two data points and gave the asymptotic normality results for the SIR estimator. Following their work, Zhu and Ng (1995) derived the asymptotic properties of the sliced estimator for general cases. Zhu and Fang (1996) proposed another version of SIR based on the kernel technique and obtained its asymptotic results. All the results summarized above are constricted to the fixed p context. Zhu, Miao, and Peng (2006) studied the asymptotic behaviors of SIR for the case where p diverges with n. A recent work (Lin, Zhao, and Liu 2017) studied the asymptotic performance of the SIR estimator from a different angle. Furthermore, SIR has been extended to the functional data and stochastic process settings, see Ferré and Yao (2003), Ferré and Yao (2005), Hsing and Ren (2009), and Li and Hsing (2010).

Instead of SIR, other methods designed for the estimation of the central subspace have also been investigated, including but not limited to the sliced average variance estimator (SAVE) (Cook and Weisberg 1991; Cook 2000), principal Hessian directions (Li 1992; Cook 1998), parametric inverse regression (Bura and Cook 2001a,b), minimum average variance estimator (Xia et al. 2002), contour regression (Li, Zha, and Chiaromonte 2005), inverse regression estimator (Cook and Ni 2005), the hybrid methods which combined SIR and SAVE in a convex way (Zhu, Ohtaki, and Li 2006), principal fitted components (Cook 2007), directional reduction (DR) (Li and Wang 2007), likelihood acquired directions (Cook and Forzani 2009), semi-parametric dimension reduction methods (Ma and Zhu 2012), and direction estimation via distance covariance (Sheng and Yin 2013, 2016).

Due to the simplicity and computational efficiency of its algorithm, SIR has been the most widely used method in practice and the most studied method in the literature. However, SIR may perform much worse when the distribution of $\mathbf{X}$ deviates from the normal case, and the SIR estimator may be inconsistent and misleading, especially in high-dimensional setting. On the one hand, this phenomenon can be seen quite clearly from our simulations below. It seems that the more the covariates $\mathbf{X}$ deviate from the multivariate normal distribution, the worse the performance of SIR gets. On the other hand, in principal component analysis (PCA), which is an unsupervised version of dimension reduction and is an intermediate step of SIR, the

CONTACT Jia Zhang zhangjia@swufe.edu.cn School of Statistics, Southwestern University of Finance and Economics, Chengdu 611130, China.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/UBES.

deviation from normality assumption may lead to the PCAs inconsistency (Johnstone and Lu 2009; Han and Liu 2016) when the dimension p of $\mathbf{X}$ is growing with the sample size n. Aware of this inconsistency problem, Han and Liu (2016) proposed a new version of PCA based on the multivariate Kendall's tau matrix for elliptically distributed covariates, called Elliptical Component Analysis (ECA). They proved that the ECA method is consistent in both sparse and non-sparse settings. In this article, we novelly extend their idea from unsupervised learning to supervised learning via a generalized eigenvalue problem (Li 2007; Chen, Zou, and Cook 2010). Consequently, our method can address the problem of low efficiency and possible inconsistency of SIR in nonnormal settings. Furthermore, our method is theoretically sound since the elliptical distribution family naturally satisfies the so-called linearity condition (Li 1991), and the merits of the introduction of the multivariate Kendall's tau matrix for elliptically distributed covariates are then well kept in the process of sufficient dimension reduction.

The main reason why we are concerned on the elliptical family is the wide applications of the elliptically distributed data, especially in finance and economics where the distribution of the data often has high kurtosis and heavy-tailed pattern. For example, Han and Liu (2016) studied a high-dimensional non-Gaussian heavy-tailed dataset on functional magnetic resonance imaging in their article. Fan, Liu, and Wang (2015) considered the problem of covariance matrix estimation based on a large factor model for elliptical data. The proof of the consistency of SIR in Lin, Zhao, and Liu (2017) was based on the assumption that the covariates $\mathbf{X}$ follows a sub-Gaussian distribution. In this article, we go a step further to investigate the elliptical family of covariates. To tackle the heavy-tailed problem, we propose a new SIR method-called elliptical sliced inverse regression (ESIR), and study it's both basic properties and high-dimensional properties. It is noteworthy that although our focus in this article is on the elliptical distribution, the applicability of the proposed method is not limited to elliptical distributed covariates, which can be seen clearly from our simulation results.

Notice that Li (1991) had a remark for the robust versions of SIR (Remark 4.4), where the author suggested the influential design points be down-weighted or be screened out in the observational study. However, things are different in our article where the focus is on the elliptical distributed covariates with heavy tails. It is not a problem of experimental design, because the data points are not under control. Besides, screening out those "bad" points seems not appropriate. On one hand, the number of those "bad" points can be very large due to the heavy tails of the covariates and removing them from the sample would worsen the estimation efficiency. On the other hand, heavy tails of the data are exactly what we care about, especially in finance and economics, and ignoring this feature might lead to misleading conclusion. To sum up, we believe that it is of great importance to do some in-deep research to address the heavy-tail related issue.

We construct the consistency and convergence rate of the ESIR estimator under the high-dimensional setting. Specifically, we allow the dimension of the covariates p, the number of the slices H and the number of the data points l in each slice to grow with the sample size n at some proper rate. This kind of study is of vital importance due to the escalating of computing power which brings us a large quantity of high-dimensional datasets in various fields, as pointed out by Zhu, Miao, and Peng (2006).

The rest of the article is organized as follows. In the next section, background knowledge is given on the elliptical distribution and the multivariate Kendall's tau matrix. In Section 3, we propose the ESIR estimator, study its basic properties and come up with an ESIR algorithm. Consistency and convergence rate of the ESIR estimator for high-dimensional covariates are investigated in Section 4. Some issues on dimension are described in Section 5. We present a large number of simulation results to compare the estimation efficiency of ESIR with that of SIR and some existing robust SIR methods, and to investigate the influence of p, H and n on the estimation accuracy in Section 6. The Istanbul stock exchange dataset is investigated in Section 7. Section 8 concludes the article and the last section reports the technical proofs.

2. Background

2.1. Elliptical Distribution

Let $\boldsymbol{\mu} \in \mathbb{R}^p$, $\mathbf{A} \in \mathbb{R}^{p\times p}$ be a deterministic matrix, $\mathbf{U} \in \mathbb{R}^p$ a uniform random vector on the unit sphere and ξ a nonnegative scaler random variable independent of $\mathbf{U}$. If

$$\mathbf{X} \overset{\mathrm{d}}{=} \boldsymbol{\mu} + \xi \mathbf{A}\mathbf{U},$$

then $\mathbf{X}$ follows an elliptical distribution, that is, $\mathbf{X} \sim EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \xi)$, where $\mathbf{A}\mathbf{A}^{\mathrm{T}} = \boldsymbol{\Sigma}$. Here, $\mathbf{X} \overset{\mathrm{d}}{=} \mathbf{Y}$ means that the random vectors $\mathbf{X}$ and $\mathbf{Y}$ follow the same distribution. Without loss of generality we assume $\mathbb{E}(\xi^2) = p$ to guarantee that $\mathrm{cov}(\mathbf{X}) = \boldsymbol{\Sigma}$.

The elliptical distribution enjoys several nice properties. That is, the marginal and conditional distributions of an elliptical distribution still belong to the elliptical family, and the independent sum of elliptical distributions is also elliptically distributed. Special cases of elliptical distribution include multivariate normal distribution, multivariate t-distribution, symmetric multivariate stable distribution, symmetric multivariate Laplace distribution and multivariate logistic distribution, etc.

Compared with the Gaussian or sub-Gaussian family, the elliptical family enables us to model complex data more flexibly. First, the elliptical family includes kinds of heavy-tailed distributions, while the Gaussian is characterized with exponential tail bounds. What is more, we can use the elliptical distribution to describe tail dependence between variables (Hult and Lindskog 2002; Han and Liu 2016). Therefore, elliptical family can be used to model complex datasets such as the financial data (Rachev 2003; Cizek 2005), genomic data (Liu 2003; Posekany 2011), and bio-imaging data (Ruttimann 1998) and so on.

2.2. Multivariate Kendall's Tau

Let $\widetilde{\mathbf{X}}$ be an independent copy of a random vector $\mathbf{X} \sim EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \xi)$. We introduce the population multivariate Kendall's tau matrix $\mathbf{M} \in \mathbb{R}^{p\times p}$ (Choi and Marden 1998)

$$\mathbf{M} := \mathbb{E}\left\{ \frac{(\mathbf{X} - \widetilde{\mathbf{X}})(\mathbf{X} - \widetilde{\mathbf{X}})^{\mathrm{T}}}{\|\mathbf{X} - \widetilde{\mathbf{X}}\|_2^2} \right\}.$$

Let $\{\mathbf{X}_i\}_{i=1}^n$ be n independent replicates of $\mathbf{X}$. The sample version of the multivariate Kendall's tau matrix is defined as

$$\widehat{\mathbf{M}} := \frac{2}{n(n-1)} \sum_{i'<i} \frac{(\mathbf{X}_i - \mathbf{X}_{i'})(\mathbf{X}_i - \mathbf{X}_{i'})^{\mathrm{T}}}{\|\mathbf{X}_i - \mathbf{X}_{i'}\|_2^2}.$$

It is straightforward to derive that $\mathbb{E}(\widehat{\mathbf{M}}) = \mathbf{M}$, $\mathrm{tr}(\widehat{\mathbf{M}}) = \mathrm{tr}(\mathbf{M}) = 1$, and $\widehat{\mathbf{M}}$ and $\mathbf{M}$ are both semi-positive definite. The sample multivariate Kendall's tau matrix is a second-order U-statistic with nice properties. Notice that the spectral norm of the kernel of the U-statistic

$$k(\mathbf{X}_i, \mathbf{X}_{i'}) := \frac{(\mathbf{X}_i - \mathbf{X}_{i'})(\mathbf{X}_i - \mathbf{X}_{i'})^{\mathrm{T}}}{\|\mathbf{X}_i - \mathbf{X}_{i'}\|_2^2}$$

is bounded by 1, which enables $\widehat{\mathbf{M}}$ to enjoy several nice theoretical properties. Furthermore, the convergence of $\widehat{\mathbf{M}}$ to $\mathbf{M}$ does not depend upon the generating variable ξ thanks to the distribution free property of the kernel (Han and Liu 2016).

Although the multivariate Kendall's tau matrix is not identical or proportional to the covariance matrix $\boldsymbol{\Sigma}$ of $\mathbf{X}$, under some mild conditions they share the same eigenspace, see Marden (1999), Croux, Ollila, and Oja (2002), Oja (2010), and Han and Liu (2016). Moreover, by simple calculation we find that $\widehat{\mathbf{M}}$ can be seen as a weighted version of the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$, that is

$$\begin{aligned}\widehat{\mathbf{M}} &= \frac{1}{n(n-1)} \sum_{i'<i} \frac{2}{\|\mathbf{X}_i - \mathbf{X}_{i'}\|_2^2} (\mathbf{X}_i - \mathbf{X}_{i'})(\mathbf{X}_i - \mathbf{X}_{i'})^{\mathrm{T}} \\ &:= \frac{1}{n(n-1)} \sum_{i'<i} \omega_{ii'} (\mathbf{X}_i - \mathbf{X}_{i'})(\mathbf{X}_i - \mathbf{X}_{i'})^{\mathrm{T}},\end{aligned}$$

while

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n(n-1)} \sum_{i'<i} (\mathbf{X}_i - \mathbf{X}_{i'})(\mathbf{X}_i - \mathbf{X}_{i'})^{\mathrm{T}}.$$

Notice that the weight is the reciprocal of the L_2 distance between $\mathbf{X}_i$ and $\mathbf{X}_{i'}$.

The multivariate Kendall's tau matrix has a kind of "shrinkage" property. For example, assuming

$$\mathbf{X} \sim N(0, \mathbf{I}_p),$$

the denominator in $\mathbf{M}$, that is $\|\mathbf{X} - \widetilde{\mathbf{X}}\|_2$, is of the order p, and $\mathbf{M}$ hence has a spectrum that is of the order $O(1/p)$. For general cases where $\mathbf{X} \sim EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \xi)$, Theorem 3.2 in Han and Liu (2016) shows that when some mild conditions hold, the jth eigenvalue of $\mathbf{M}$ satisfies $\lambda_j(\mathbf{M}) \asymp \lambda_j(\boldsymbol{\Sigma})/\mathrm{trace}(\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ denotes the covariance matrix of $\mathbf{X}$. Furthermore, when the condition number of $\boldsymbol{\Sigma}$ is upper bounded by an absolute constant, we can obtain $\lambda_j(\mathbf{M}) \asymp \lambda_j(\boldsymbol{\Sigma})/(\|\boldsymbol{\Sigma}\|_{\mathrm{F}}\sqrt{p})$, which means that the spectrum of $\boldsymbol{\Sigma}$ would be of order $O(1/\sqrt{p})$ in general cases. Due to this shrinking to zero property of the Kendall's tau matrix, it makes little sense to compare $\mathbf{M}$ and its estimate $\widehat{\mathbf{M}}$.

A useful and common fix is to redefine $\mathbf{M}$ as

$$\mathbf{M} := \mathbb{E}\left\{ \frac{(\mathbf{X} - \widetilde{\mathbf{X}})(\mathbf{X} - \widetilde{\mathbf{X}})^{\mathrm{T}}}{\|\mathbf{X} - \widetilde{\mathbf{X}}\|_2^2/p} \right\}. \tag{2.1}$$

Then, its sample estimator becomes

$$\widehat{\mathbf{M}} := \frac{2}{n(n-1)} \sum_{i'<i} \frac{(\mathbf{X}_i - \mathbf{X}_{i'})(\mathbf{X}_i - \mathbf{X}_{i'})^{\mathrm{T}}}{\|\mathbf{X}_i - \mathbf{X}_{i'}\|_2^2/p}.$$

The newly defined $\mathbf{M}$ and $\widehat{\mathbf{M}}$ will be used in the following analysis.

3. Elliptical Sliced Inverse Regression

3.1. Sliced Inverse Regression

In this section, we give a brief introduction of the SIR method. The model below is considered

$$Y = f(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}, \ldots, \boldsymbol{\beta}_K^{\mathrm{T}}\mathbf{X}, \boldsymbol{\epsilon}), \tag{3.1}$$

where $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ are unknown p dimensional column vectors, $\boldsymbol{\epsilon}$ is independent of the covariates $\mathbf{X}$, and f is an arbitrary unknown function defined on $\mathbb{R}^{K+1}$. The linear space $\mathcal{B}$ generated by $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ is called the efficient dimension reduction (e.d.r.) space, and any linear combination of $\boldsymbol{\beta}$'s is referred to as an e.d.r. direction.

Li (1991) demonstrated that if $\mathbf{X}$ was standardized by $\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{X})$ to have zero mean and identity covariance matrix, the inverse regression curve $\mathbb{E}(\mathbf{X}|Y)$ would be contained in the e.d.r. space. Accordingly, the PCA method can be applied to the estimated covariance matrix of the inverse regression curve. Hence, the leading eigenvectors of the estimated covariance matrix can then be transformed to estimate the e.d.r. directions. Furthermore, Li (1991) showed that each estimator $\hat{\boldsymbol{\beta}}_k$ converges to an e.d.r. direction at rate of $n^{-1/2}$ when p stays fixed. The essential condition for the SIR method is referred to as the linearity condition, that is, for any $\mathbf{b} \in \mathbb{R}^p$, $\mathbb{E}(\mathbf{b}^{\mathrm{T}}\mathbf{X}|\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}, \ldots, \boldsymbol{\beta}_K^{\mathrm{T}}\mathbf{X}) = c_0 + c_1\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X} + \ldots + c_K\boldsymbol{\beta}_K^{\mathrm{T}}\mathbf{X}$ for some constants $c_0, \ldots, c_K$. This condition requires that the distribution of the covariates be elliptically symmetric. Such distributions include the normal distribution and the general elliptical distributions.

3.2. Elliptical Sliced Inverse Regression

We construct a basic theorem for ESIR in this part. Here, "E" represents our focus on the elliptical family which is characterized with heavy tails.

Theorem 1. Assume that $\mathbf{X} \sim EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \xi)$. Under (3.1), the curve $\mathbb{E}(\mathbf{X}|Y) - \mathbb{E}(\mathbf{X})$ is contained in the linear subspace spanned by $\mathbf{M}\boldsymbol{\beta}_k (k = 1, 2, \ldots, K)$, where $\mathbf{M}$ denotes the multivariate Kendall's tau matrix of $\mathbf{X}$.

Theorem 1 seems to be similar to Theorem 3.1 of Li (1991). It is not surprising in view of the close relationship between $\boldsymbol{\Sigma}$ and $\mathbf{M}$ in Section 2. It is worth noting that we only assume that the covariates $\mathbf{X}$ follow an elliptical distribution and do not pose any distribution restriction on $\mathbf{X}|Y$, while Bura and Forzani (2015) and Bura, Duarte, and Forzani (2016) required $\mathbf{X}|Y$ be elliptically distributed and multivariate exponentially distributed, respectively.

Let $\mathbf{Z} = \mathbf{M}^{-1/2}\{\mathbf{X} - \mathbb{E}(\mathbf{X})\}$, where $\mathbf{M}$ is the population multivariate Kendall's tau matrix defined above in (2.1). Then, (3.1) can be rewritten as

$$Y = f(\boldsymbol{\eta}_1^{\mathrm{T}}\mathbf{Z}, \ldots, \boldsymbol{\eta}_K^{\mathrm{T}}\mathbf{Z}, \boldsymbol{\epsilon}), \tag{3.2}$$

where $\boldsymbol{\eta}_k = \mathbf{M}^{1/2}\boldsymbol{\beta}_k (k = 1, \ldots, K)$. Following the usage in Li (1991), the vector linearly generated by the $\boldsymbol{\eta}_k$'s is called a standardized e.d.r direction. For this new version of standardized covariates, we have the following corollary.

Corollary 1. Under (3.2), the curve $\mathbb{E}(\mathbf{Z}|Y)$ is contained in the linear space generated by $\boldsymbol{\eta}_k$'s defined in Equation (3.2).

Corollary 1 implies that $\mathrm{cov}\{\mathbb{E}(\mathbf{Z}|Y)\}$ is degenerate in directions orthogonal to $\boldsymbol{\eta}_k$'s. The following proposition gives the relationship between $\mathrm{cov}\{\mathbb{E}(\mathbf{X}|Y)\}$ and $\mathbf{M}_{\mathbb{E}(\mathbf{X}|Y)}$, the population multivariate Kendall's tau matrix of the inverse regression curve $\mathbb{E}(\mathbf{X}|Y)$.

Proposition 1. Assume $\mathbf{X} \sim EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \xi)$. Write $\mathrm{cov}\{\mathbb{E}(\mathbf{X}|Y)\} = \boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Omega}^{\mathrm{T}}$, where $\boldsymbol{\Omega} = (\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(p)})^{\mathrm{T}}$ is the $p \times p$ matrix of the eigenvectors and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ being the corresponding eigenvalues. Letting $\mathbf{M}_{\mathbb{E}(\mathbf{X}|Y)}$ denote the population multivariate Kendall's tau matrix of the vector $\mathbb{E}(\mathbf{X}|Y)$, it holds that

$$\mathbf{M}_{\mathbb{E}(\mathbf{X}|Y)} = \boldsymbol{\Omega}(p\boldsymbol{\Lambda}_1)\boldsymbol{\Omega}^{\mathrm{T}},$$

where $\boldsymbol{\Lambda}_1$ is a $p \times p$ diagonal matrix containing the eigenvalues of $\mathbf{M}_{\mathbb{E}(\mathbf{X}|Y)}$.

Notice that the key point in SIR is to estimate the leading eigenvectors of some covariance matrix, and the corresponding eigenvalues can be treated as nuisance parameters. By Proposition 1, we can conclude that

$$\mathbf{M}_{\mathbb{E}(\mathbf{Z}|Y)} = \mathbb{E}\left[\frac{\{\mathbb{E}(\mathbf{Z}|Y) - \mathbb{E}(\widetilde{\mathbf{Z}}|\widetilde{Y})\}\{\mathbb{E}(\mathbf{Z}|Y) - \mathbb{E}(\widetilde{\mathbf{Z}}|\widetilde{Y})\}^{\mathrm{T}}}{\|(\mathbb{E}(\mathbf{Z}|Y) - \mathbb{E}(\widetilde{\mathbf{Z}}|\widetilde{Y}))\|_2^2/p}\right]$$

is also degenerate in directions orthogonal to $\boldsymbol{\eta}_k$'s. Thus, the eigenvectors associated with the largest K eigenvalues of $\mathbf{M}_{\mathbb{E}(\mathbf{Z}|Y)}$ are the standardized e.d.r. directions $\boldsymbol{\eta}_k (k = 1, \ldots, K)$. We then transform $\boldsymbol{\eta}_k (k = 1, \ldots, K)$ to the original e.d.r. direction $\boldsymbol{\beta}_k$ by $\boldsymbol{\beta}_k = \mathbf{M}^{-1/2}\boldsymbol{\eta}_k (k = 1, \ldots, K)$.

Given Corollary 1 and Proposition 1, we construct the operating scheme for the ESIR:

---

1. For each $\mathbf{X}_i (i = 1, 2, \ldots, n)$, calculate its standardized form: $\mathbf{Z}_i = \widehat{\mathbf{M}}^{-1/2}(\mathbf{X}_i - \bar{\mathbf{X}})(i = 1, 2, \ldots, n)$, where $\widehat{\mathbf{M}}$ and $\bar{\mathbf{X}}$ denote the sample multivariate Kendall's tau matrix and the sample mean of $\mathbf{X}$, respectively.
2. Divide the range of Y into H "equal" slices, $I_1, \ldots, I_H$, where "equal" means that the number of the data points falling in each slice is equal to $l = \lfloor n/H \rfloor$.
3. In each slice, compute the sample mean of $\mathbf{Z}$: $\hat{\mathbf{m}}_h = 1/l \sum_{Y_i \in I_h} \mathbf{Z}_i$, where $h = 1, \ldots, H$.
4. Compute the multivariate Kendall's tau matrix for $\hat{\mathbf{m}}_h$:
$$\widehat{\mathbf{M}}_{\mathbf{m}} = \frac{2}{H(H-1)} \sum_{h' < h} \frac{(\hat{\mathbf{m}}_h - \hat{\mathbf{m}}_{h'})(\hat{\mathbf{m}}_h - \hat{\mathbf{m}}_{h'})^{\mathrm{T}}}{\|\hat{\mathbf{m}}_h - \hat{\mathbf{m}}_{h'}\|_2^2/p} \quad (h = 1, \ldots, H). \tag{3.3}$$
Then find the eigenvalues and eigenvectors of $\widehat{\mathbf{M}}_{\mathbf{m}}$.
5. Denote the top K leading eigenvectors of $\widehat{\mathbf{M}}_{\mathbf{m}}$ be $\hat{\boldsymbol{\eta}}_k (k = 1, \ldots, K)$. Transform them back to the original e.d.r. directions by $\widehat{\mathbf{M}}$, that is, $\hat{\boldsymbol{\beta}}_k = \widehat{\mathbf{M}}^{-1/2}\hat{\boldsymbol{\eta}}_k$.

In this algorithm, the number of the data points in each slice is enforced to be fixed to l so that we do not need to do any weighting adjustment for the calculation of $\widehat{\mathbf{M}}_{\mathbf{m}}$. In practice, the data points in the last slice may not be exactly l, which exerts little influence on the estimation asymptotically. Our algorithm actually belongs to a generalized eigenvalue framework. Please see more details in Li (2007) and Chen, Zou, and Cook (2010). It is remarkable that, like SIR, ESIR may not recover all the e.d.r. directions. One may refer to other dimension reduction methods like SAVE and DR to address such problems.

Remark 1. In Steps 1 and 3, sample mean is used for centralization. For robustness purpose, robust mean estimators, like coordinate mean, spatial mean, or median-of-means, can be used to approximate the population mean. Simulation results show that using robust mean estimators in the above ESIR algorithm improves estimation efficiency. See the simulation results in the supplementary materials.

4. Asymptotic Properties of Elliptical Slice Inverse Regression With Diverging Number of Covariates

In this article, we assume the number l of the data points in each slice stays the same, and the number of the slices H and l are both allowed to grow with the sample size n. In the proof, we use the original covariates $\mathbf{X}$ rather than its standardized version for simplicity. The conclusion can be directly extended to the standardized version.

Denote the inverse regression curve by $\mathbf{m}(Y) = \mathbb{E}(\mathbf{X}|Y)$ and decompose $\mathbf{X}$ as:

$$\mathbf{X} = \mathbf{m}(Y) + \boldsymbol{\varepsilon},$$

where $\mathbf{m}(Y) = \{m_1(Y), \ldots, m_p(Y)\}^{\mathrm{T}}$ with $m_i(Y) = \mathbb{E}(X_i|Y)$ and $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$. For the sample version, let

$$\mathbf{X}_i = \mathbf{m}(Y_i) + \boldsymbol{\varepsilon}_i = \mathbf{m}_i + \boldsymbol{\varepsilon}_i, \qquad i = 1, \ldots, n$$

and

$$\mathbf{X}_{(i)} = \mathbf{m}(Y_{(i)}) + \boldsymbol{\varepsilon}_{(i)} = \mathbf{m}_{(i)} + \boldsymbol{\varepsilon}_{(i)}, \qquad i = 1, \ldots, n$$

where $Y_{(1)} \leq \ldots, \leq Y_{(n)}$ and $\mathbf{X}_{(i)}$ and $\boldsymbol{\varepsilon}_{(i)}$ are the concomitants (Yang 1977) of $Y_{(i)}$. For each slice, denote

$$\mathbf{X}_{hi} = \mathbf{m}(Y_{hi}) + \boldsymbol{\varepsilon}_{hi} = \mathbf{m}_{hi} + \boldsymbol{\varepsilon}_{hi}, \quad i = 1, \ldots, l, \quad h = 1, \ldots, H,$$

where $\mathbf{X}_{hi} = \mathbf{X}_{l(h-1)+i}$ and $Y_{hi} = Y_{l(h-1)+i}$. The following conditions are need.

Condition 1. $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}} \sim EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \xi)$ and $\sup_{1 \leq j \leq p} \mathbb{E}(|X_j|^m) < \infty$ for some constant $m \geq 2$.

Condition 2. There exists a positive constant C such that $\lambda_{\max}\{\mathbf{M}_{\mathbb{E}(\mathbf{X}|Y)}\} \leq C$.

Condition 3. Let $\mathbf{m}_h = \mathbb{E}(\mathbf{X}|Y \in I_h)$ for $h = 1, \ldots, H$, and assume

$$\max_{h \neq h'} \frac{1}{\|\mathbf{m}_h - \mathbf{m}_{h'}\|_2} < \infty, \quad \max_{h \neq h', j, k} \frac{\|\mathbf{m}_{hj} - \mathbf{m}_{h'k}\|_2}{\|\mathbf{m}_h - \mathbf{m}_{h'}\|_2} < \infty,$$
$$\max_{h,j,k} |Y_{hk} - Y_{hj}| = O_p(1/H).$$

The second part of Condition 1 is similar to that of Hsing and Carroll (1992), Zhu and Ng (1995), and Zhu, Miao, and Peng (2006), which require $m \geq 4$ rather than $m \geq 2$ in our work. The reason why we have a much milder condition may originate from the first part of this condition, that is, the covariates $\mathbf{X}$ is restricted to be elliptically distributed. Condition 2 seems to be a quite mild condition, which is reasonable in view of the connection between $\mathbb{E}(\mathbf{X}|Y)$ and $\mathbf{X}$. The last condition can be subtly related to the "ϑ-stable" in Condition 3 in Lin, Zhao, and Liu (2017).

Notice that although Han and Liu (2016) required $\mathbf{X} \sim EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \xi)$ in their Theorem 3.1, one can easily find that it is not a necessary condition by reviewing their proof carefully, which means that we do not need any restriction on the distribution of $\mathbb{E}(\mathbf{X}|Y)$ for the consistency of the estimator in our article. This is a quite appealing property, because it is not trivial to test the distribution of $\mathbb{E}(\mathbf{X}|Y)$.

Under the above conditions, we establish the consistency and convergence rate of the ESIR estimator in the following theorems. Recall that $\widehat{\mathbf{M}}_m$ is defined in Equation (3.3) and $\mathbf{M}_{\mathbb{E}(\mathbf{X}|Y)}$ denotes the population Kendall's tau matrix of $\mathbb{E}(\mathbf{X}|Y)$. We first introduce a useful proposition.

Proposition 2. Under Conditions 1-3, if $p = o(n^{1/2})$, then we have

$$\|\widehat{\mathbf{M}}_m - \mathbf{M}_{\mathbb{E}(\mathbf{X}|Y)}\|_2 = O_p(p^{3/2}H^{-1}) + O_p(pH^{1/2}n^{-1/2}).$$

Proposition 2 shows that the effect of H on the convergence rate is two-sided. Hence, we suggest choosing a moderate size for the number of slices, not too big nor too small. On the other hand, by choosing a proper diverging speed of the number H of the slices, we can obtain $\|\widehat{\mathbf{M}}_\mathbf{m} - \mathbf{M}_{\mathbb{E}(\mathbf{X}|Y)}\|_2 = O_p(pH^{1/2}n^{-1/2})$.

Denote the top K leading eigenvectors of $\mathbf{M}_{\mathbb{E}(\mathbf{X}|Y)}$ by $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K$ and the top K leading eigenvectors of the slice-based estimator $\widehat{\mathbf{M}}_m$ of $\mathbf{M}_{\mathbb{E}(\mathbf{X}|Y)}$ by $\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_K$. Then, by Theorem 1, $\boldsymbol{\beta}_k$ in Model (3.1) can be estimated by $\hat{\boldsymbol{\beta}}_k = \widehat{\mathbf{M}}^{-1}\hat{\boldsymbol{\gamma}}_k$ for $k = 1, \ldots, K$. Under Model (3.1), the $p \times K$ matrix $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$ is not identifiable, but the space spanned by the $\boldsymbol{\beta}_k$'s can be. Hence, in the following, we consider the convergence of the column space of $\widehat{\mathbf{B}} = (\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_K)$ to the column space of $\mathbf{B}$. To this end, the following theorem is of key importance.

Theorem 2. Under Conditions 1–3, if the eigenvalues of $\mathbf{M}$ are bounded away from zero and infinity, $p^{3/2}H^{-1} = o(1)$ and $pH^{1/2}n^{-1/2} = o(1)$, it holds that

$$\|\widehat{\mathbf{M}}^{-1}\widehat{\mathbf{M}}_\mathbf{m} - \mathbf{M}^{-1}\mathbf{M}_{\mathbb{E}(\mathbf{X}|Y)}\|_2 = O_p(p^{3/2}H^{-1}) + O_p(pH^{1/2}n^{-1/2}).$$

Let $P_\mathbf{B}$ be the projection matrix onto the column space of $\mathbf{B}$, and $P_{\widehat{\mathbf{B}}}$ be similarly defined. The following corollary can be obtained by Theorem 2 and a variation of the Devis-kahan inequality (Vu and Lei 2013).

Corollary 2. Under the conditions of Theorem 2, it holds that

$$\|P_{\widehat{\mathbf{B}}} - P_\mathbf{B}\|_2 \xrightarrow{p} 0$$

at the same rate as that in Theorem 2.

Specifically, when $K = 1$, the convergence of $\hat{\boldsymbol{\gamma}}_1$ to $\boldsymbol{\gamma}_1$ can be controlled by the spectral norm of $\widehat{\mathbf{M}}_\mathbf{m} - \mathbf{M}_{\mathbb{E}(\mathbf{X}|Y)}$ by the Davis–Kahan inequality (Davis and Kahan 1970; Wedin 1972). In detail, the Davis–Kahan inequality states that

$$|\sin\angle(\hat{\boldsymbol{\gamma}}_1, \boldsymbol{\gamma}_1)| \le \frac{2}{\lambda_{\max}\{\mathbf{M}_{\mathbb{E}(\mathbf{X}|Y)}\}}\|\widehat{\mathbf{M}}_\mathbf{m} - \mathbf{M}_{\mathbb{E}(\mathbf{X}|Y)}\|_2,$$

where $|\sin\angle(\mathbf{v}_1, \mathbf{v}_2)| = \sqrt{1 - (\mathbf{v}_1^\mathrm{T}\mathbf{v}_2)^2}$ for any two vectors $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^p$. Notice that $\|\widehat{\mathbf{M}} - \mathbf{M}\|_2 = O_p\{p^{1/2}(\log p)^{1/2}n^{-1/2}\}$, and then we obtain

$$|\sin\angle(\hat{\boldsymbol{\beta}}_1, \boldsymbol{\beta}_1)| \xrightarrow{p} 0 \quad \text{as} \quad n \to \infty,$$

provided that $p^{3/2}H^{-1} = o(1)$ and $pH^{1/2}n^{-1/2} = o(1)$.

5. Dimension Issues

For convenience of theoretical derivation, we assume that the dimension K of the central subspace is known. In practice, some criterion or tests are needed to determine K. We do not pay much attention to this problem in this article because there have been several methods in the literature which can be employed in the elliptical setting. For example, Schott (1994) extended the sequential chi-squared test procedure of Li (1991) to the setting of elliptically distributed covariates, Bura and Cook (2001a) proposed a general weighted chi-squared sequential test, Zhu, Miao, and Peng (2006) suggested a Bayes information criterion type procedure to ascertain the dimension of the central subspace and Chen, Cook, and Zou (2015) proposed a test based on conditional distance covariance to check the goodness-of-fit of a given dimension reduction subspace.

From Theorem 2, it is easy to find that although the dimension p is allowed to diverge with the sample size n, we still require $p \ll n$. That is because we need to invert the multivariate Kendall's tau matrix in the algorithm. When p is larger than n, especially in the ultrahigh-dimensional setting where $\log p = O(n^\alpha)$ for some $0 < \alpha < 1$, we can follow the framework of two-scale statistical learning (Fan and Lv 2008) by doing large-scale screening first followed by moderate-scale sufficient dimension reduction. Model free feature screening methods like screening via distance correlation (Li 2012), the method proposed by Zhu et al. (2011) and screening via Ball correlation (BCor-SIS) (Pan et al. 2018) can be exploited. We suggest BCor-SIS be employed because it is robust to heavy-tailed variables. After the screening procedure, the dimension can be reduced to a scale of $O\{n/\log(n)\}$, then sufficient dimension reduction methods like ESIR can be readily applied in the reduced feature space. If the sure screening property (Fan and Lv 2008) holds in the screening stage, the advantage of the ESIR method can be retained in the ultrahigh-dimensional setting. This two-scale learning framework is in the spirit of Fan and Lv (2008) for sure independence screening.

6. Numerical Examples

A variety of numerical examples are reported in this section. The squared multiple correlation coefficient $R^2(\hat{\boldsymbol{\beta}}_i)$ (Li 1991; Zhu, Miao, and Peng 2006) is used to measure the distance between the ESIR estimator $\hat{\boldsymbol{\beta}}_i$ and the central subspace $\mathcal{B}$ for $i = 1, \ldots, K$ and their average R^2 to measure the distance between the space formed by all the $\hat{\boldsymbol{\beta}}$'s and the central subspace (Li 1991; Zhu, Miao, and Peng 2006). For any $p \times 1$ vector $\mathbf{b}$, $R^2(\mathbf{b})$ is calculated by

$$R^2(\mathbf{b}) = \max_{\boldsymbol{\beta} \in \mathcal{B}} \frac{(\mathbf{b}^\mathrm{T}\boldsymbol{\Sigma}\boldsymbol{\beta})^2}{\mathbf{b}^\mathrm{T}\boldsymbol{\Sigma}\mathbf{b} \cdot \boldsymbol{\beta}^\mathrm{T}\boldsymbol{\Sigma}\boldsymbol{\beta}}. \quad (6.1)$$

Thus, a bigger squared multiple correlation coefficient indicates higher estimation efficiency.

6.1. Single-Index Models

Three types of single-index models are considered under multivariate normal distribution and other five frequently used elliptical distributions in this part, including the multivariate Laplace distribution, multivariate symmetric logistic distribution, multivariate Student's t distribution with degrees of freedom 2 and 3 and the multivariate Cauchy distribution. Notice that although the multivariate Cauchy distribution does not satisfy Condition 1, the ESIR method performs surprisingly well in the models given below.

Model (A1):

$$Y = \frac{1}{0.5 + (\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X} + 1.5)^2} + \sigma\epsilon.$$

Model (A2):

$$Y = 0.5 + (\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X} + 1.5)^2 + \sigma\epsilon.$$

Model (A3):

$$Y = (\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X} + 2) \cdot \sigma\epsilon.$$

Model (A1) and (A2) come from Li (1991), and Model (A3) is motivated by Example 3 of Zhu, Miao, and Peng (2006). In all the above three models, $\sigma = 0.5$, $\boldsymbol{\beta}_1 = (1, 0, \ldots, 0)^{\mathrm{T}}$, $\epsilon \sim N(0, 1)$ and $\mathbf{X} \sim EC_p(\mathbf{0}, \boldsymbol{\Sigma}, \xi)$ where $\boldsymbol{\Sigma} = \mathbf{I}_{p\times p}$ and ξ is the generating variable. We change the distribution of $\mathbf{X}$ among the elliptical distributions mentioned above by altering the distribution of the generating variable ξ. For multivariate logistic distribution, the dependence parameter is chosen to be 0.2 to produce weak dependence among the elements of $\mathbf{X}$. The sample size n, the number of predictors p and the number of the slices H are chosen to be 400, 10, and 10, respectively. Table 1 reports the means and standard deviations of $R^2(\hat{\boldsymbol{\beta}}_1)$ after 100 replicates under different simulation schemes.

Table 1. Mean and standard deviation (in parentheses) of $R^2(\hat{\boldsymbol{\beta}}_1)$ for the single-index models.

Distr of X	Normal	Laplace	Logistic	t (3)	t (2)	Cauchy
	$R^2(\hat{\boldsymbol{\beta}}_1)$	$R^2(\hat{\boldsymbol{\beta}}_1)$	$R^2(\hat{\boldsymbol{\beta}}_1)$	$R^2(\hat{\boldsymbol{\beta}}_1)$	$R^2(\hat{\boldsymbol{\beta}}_1)$	$R^2(\hat{\boldsymbol{\beta}}_1)$
Model (A1)						
SIR	0.95	0.88	0.90	0.71	0.37	0.10
	(0.02)	(0.07)	(0.05)	(0.19)	(0.25)	(0.12)
ESIR	0.95	0.88	0.97	0.84	0.60	0.47
	(0.02)	(0.08)	(0.00)	(0.09)	(0.30)	(0.33)
Model (A2)						
SIR	1.00	0.97	1.00	0.77	0.42	0.18
	(0.00)	(0.02)	(0.00)	(0.23)	(0.28)	(0.16)
ESIR	1.00	0.98	0.97	0.93	0.81	0.48
	(0.00)	(0.01)	(0.00)	(0.05)	(0.18)	(0.36)
Model (A3)						
SIR	0.91	0.82	0.90	0.66	0.34	0.16
	(0.04)	(0.12)	(0.06)	(0.25)	(0.27)	(0.15)
ESIR	0.90	0.80	0.97	0.73	0.49	0.40
	(0.05)	(0.13)	(0.01)	(0.20)	(0.30)	(0.34)

As can be seen from Table 1, the ESIR method outperforms the SIR method in almost all the simulation schemes. Furthermore, the efficiency gain is more significant when the tail of the distribution of $\mathbf{X}$ tends to become heavier, which can be easily seen from the simulation results of $t(3)$, $t(2)$, and Cauchy ($t(1)$) distributed covariates where a smaller degree of Student's t distribution indicates a heavier tail. For multivariate normal distribution and Laplace distribution, our ESIR estimator performs nearly as well as the SIR estimator. Notice that the tail of the Laplace distribution is very close to that of the normal distribution.

6.2. Double Index Models

Four models are considered in this section with $K = 2$ for six different elliptical distributions of $\mathbf{X}$. Unless otherwise noted, the simulation parameters used here keep the same as those used in the first part for the single index models.

Model (B1):

$$Y = \frac{\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}}{0.5 + (\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{X} + 1.5)^2} + \sigma\epsilon,$$

where $\boldsymbol{\beta}_1 = (1, 0, \ldots, 0)^{\mathrm{T}}$, $\boldsymbol{\beta}_2 = (0, 1, 0, \ldots, 0)^{\mathrm{T}}$, $\epsilon \sim N(0, 1)$ and $\mathbf{X} \sim EC_p(\mathbf{0}, \boldsymbol{\Sigma}, \xi)$ where $\boldsymbol{\Sigma} = \mathbf{I}_{p\times p}$. This model was used by Li (1991).

Model (B2):

$$Y = 4 + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X} + (\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{X} + 2) \cdot \sigma\epsilon.$$

Here, we reset $p = 5$, $\mathbf{X} \sim EC_p(\mathbf{0}, \boldsymbol{\Sigma}, \xi)$ with $\boldsymbol{\Sigma} = \mathrm{diag}\{2, 2, 2, 4, 2\}$, $\boldsymbol{\beta}_1 = (1, 0, 0, 0, 0)^{\mathrm{T}}$ and $\boldsymbol{\beta}_2 = (0, 1, 1, 0, 0)^{\mathrm{T}}$.

Model (B3):

$$Y = (4 + \boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}) \cdot (\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{X} + 2) + \sigma\epsilon.$$

The simulation parameters for this model are the same as those in model (B2). Model (B2) and (B3) come from Examples 2 and 3 of Zhu, Miao, and Peng (2006) respectively.

Model (B4):

$$Y = (\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X})^2 + |\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{X}| + \sigma\epsilon,$$

where $\boldsymbol{\beta}_1 = (0.5, 0.5, 0.5, 0.5, 0, \ldots, 0)^{\mathrm{T}}$ and $\boldsymbol{\beta}_2 = (0.5, -0.5, 0.5, -0.5, 0, \ldots, 0)^{\mathrm{T}}$. This model comes from Example 3 of Chen, Cook, and Zou (2015). The distribution of $\mathbf{X}$ deviates a little bit from the elliptical distribution. That is, let $\mathbf{X} = (X_1, \mathbf{X}_2)$ where $\mathbf{X}_2 = (X_2, \ldots, X_p)$, $\mathbf{X}_2 \sim EC_{p-1}(\mathbf{0}, \boldsymbol{\Sigma}, \xi)$, where $\boldsymbol{\Sigma} = (\sigma_{ij})$ with $\sigma_{ij} = 0.5^{|i-j|}$ for $i, j = 1, \ldots, (p-1)$ and $X_1 = |X_2 + X_3| + \zeta$ with $\zeta \sim N(0, 1)$.

In these $K = 2$ cases, we remove the Laplace distribution and add another elliptical distribution $EC1$, because the result based on the Laplace distributed covariates is quite similar to that of the normal distribution. Define $EC1 = EC_p(\mathbf{0}, \boldsymbol{\Sigma}, \xi_1)$ with $\xi_1 \sim F(p, 1)$ where F indicates F distribution. Notice that ξ_1 does not have finite mean. This distribution was also used in Han and Liu (2016). Tables 2 and 3 exhibit a little bit difference from the result of single index models. That is, while the first leading eigenvector or direction presents almost the same efficiency improvement as in the single index case, the second estimated direction does not perform so well as the SIR

Table 2. Mean and standard deviation (in parentheses) of $R^2(\hat{\beta})$ for double index models I.

Distr of X	Normal $R^2(\hat{\beta}_1)$	Normal $R^2(\hat{\beta}_2)$	Normal R^2	Logistic $R^2(\hat{\beta}_1)$	Logistic $R^2(\hat{\beta}_2)$	Logistic R^2	EC1 $R^2(\hat{\beta}_1)$	EC1 $R^2(\hat{\beta}_2)$	EC1 R^2
Model (B1)									
SIR	0.96	0.88	0.92	0.91	0.20	0.56	0.22	0.19	0.21
	(0.02)	(0.06)		(0.04)	(0.18)		(0.18)	(0.15)	
ESIR	0.94	0.76	0.85	0.99	0.18	0.59	0.89	0.84	0.87
	(0.03)	(0.16)		(0.00)	(0.17)		(0.20)	(0.24)	
Model (B2)									
SIR	0.99	0.81	0.90	0.99	0.71	0.85	0.48	0.44	0.46
	(0.01)	(0.22)		(0.01)	(0.27)		(0.23)	(0.26)	
ESIR	0.99	0.62	0.81	1.00	0.74	0.87	0.94	0.85	0.90
	(0.01)	(0.31)		(0.00)	(0.24)		(0.13)	(0.24)	
Model (B3)									
SIR	1.00	0.93	0.97	1.00	0.26	0.63	0.38	0.42	0.40
	(0.00)	(0.05)		(0.00)	(0.26)		(0.21)	(0.29)	
ESIR	1.00	0.78	0.89	1.00	0.23	0.62	0.88	0.87	0.88
	(0.00)	(0.20)		(0.00)	(0.24)		(0.23)	(0.20)	
Model (B4)									
SIR	0.97	0.66	0.82	1.00	0.67	0.84	0.43	0.42	0.43
	(0.01)	(0.20)		(0.00)	(0.21)		(0.15)	(0.07)	
ESIR	0.97	0.69	0.83	1.00	0.90	0.95	0.92	0.81	0.87
	(0.01)	(0.16)		(0.00)	(0.03)		(0.17)	(0.28)	

Table 3. Mean and standard deviation (in parentheses) of $R^2(\hat{\beta})$ for double index models II.

Distr of X	$t(3)$ $R^2(\hat{\beta}_1)$	$t(3)$ $R^2(\hat{\beta}_2)$	$t(3)$ R^2	$t(2)$ $R^2(\hat{\beta}_1)$	$t(2)$ $R^2(\hat{\beta}_2)$	$t(2)$ R^2	Cauchy ($t(1)$) $R^2(\hat{\beta}_1)$	Cauchy ($t(1)$) $R^2(\hat{\beta}_2)$	Cauchy ($t(1)$) R^2
Model (B1)									
SIR	0.89	0.72	0.81	0.77	0.48	0.63	0.25	0.21	0.23
	(0.10)	(0.15)		(0.16)	(0.20)		(0.21)	(0.15)	
ESIR	0.93	0.54	0.74	0.90	0.49	0.70	0.79	0.66	0.73
	(0.05)	(0.24)		(0.07)	(0.27)		(0.26)	(0.28)	
Model (B2)									
SIR	0.98	0.42	0.70	0.90	0.36	0.63	0.67	0.33	0.50
	(0.02)	(0.32)		(0.13)	(0.27)		(0.23)	(0.22)	
ESIR	0.98	0.32	0.65	0.98	0.37	0.68	0.95	0.67	0.81
	(0.02)	(0.28)		(0.03)	(0.30)		(0.09)	(0.33)	
Model (B3)									
SIR	0.96	0.73	0.85	0.85	0.47	0.66	0.40	0.41	0.41
	(0.05)	(0.25)		(0.15)	(0.28)		(0.27)	(0.27)	
ESIR	0.99	0.59	0.79	0.95	0.47	0.71	0.86	0.68	0.77
	(0.01)	(0.30)		(0.06)	(0.31)		(0.20)	(0.28)	
Model (B4)									
SIR	0.93	0.33	0.63	0.85	0.22	0.54	0.62	0.12	0.37
	(0.07)	(0.24)		(0.14)	(0.20)		(0.19)	(0.13)	
ESIR	0.91	0.40	0.66	0.87	0.34	0.61	0.86	0.56	0.71
	(0.05)	(0.25)		(0.10)	(0.25)		(0.15)	(0.32)	

estimator under several simulation settings. However, one can find that when the tail of the distribution gets heavier, the ESIR estimation for the second e.d.r. direction inclines to become more accurate. From a comprehensive point of view, the ESIR estimation efficiency is comparable to or better than that of the SIR method. The robustness of ESIR is well demonstrated in Table 3, that is, when the tail of the distribution of the covariates goes heavier (from t(3) to t(1)), the performance of the proposed

Table 4. Estimation of the central subspace for Model (B1) with Cauchy distributed covariates.

H		5			10			20			40		
p		5	10	30	5	10	30	5	10	30	5	10	30
$n = 120$													
SIR	$R^2(\hat{\beta}_1)$	0.47	0.24	0.12	0.47	0.27	0.12	0.46	0.24	0.10	0.41	0.24	0.09
	$R^2(\hat{\beta}_2)$	0.42	0.21	0.08	0.41	0.22	0.07	0.42	0.24	0.08	0.41	0.17	0.08
ESIR	$R^2(\hat{\beta}_1)$	0.86	0.82	0.74	0.83	0.77	0.73	0.87	0.76	0.69	0.82	0.74	0.67
	$R^2(\hat{\beta}_2)$	0.62	0.60	0.56	0.68	0.60	0.59	0.70	0.66	0.54	0.75	0.64	0.58
$n = 200$													
SIR	$R^2(\hat{\beta}_1)$	0.52	0.29	0.13	0.44	0.28	0.13	0.48	0.32	0.10	0.43	0.25	0.10
	$R^2(\hat{\beta}_2)$	0.37	0.25	0.08	0.44	0.21	0.09	0.43	0.21	0.07	0.40	0.17	0.09
ESIR	$R^2(\hat{\beta}_1)$	0.87	0.78	0.70	0.86	0.75	0.63	0.85	0.79	0.72	0.87	0.74	0.68
	$R^2(\hat{\beta}_2)$	0.66	0.56	0.49	0.67	0.58	0.57	0.66	0.62	0.54	0.65	0.66	0.56
$n = 400$													
SIR	$R^2(\hat{\beta}_1)$	0.48	0.29	0.11	0.48	0.28	0.10	0.48	0.26	0.12	0.43	0.25	0.09
	$R^2(\hat{\beta}_2)$	0.40	0.21	0.08	0.42	0.22	0.06	0.39	0.19	0.07	0.39	0.20	0.07
ESIR	$R^2(\hat{\beta}_1)$	0.87	0.80	0.76	0.91	0.73	0.72	0.92	0.86	0.67	0.88	0.77	0.67
	$R^2(\hat{\beta}_2)$	0.68	0.58	0.56	0.70	0.60	0.64	0.71	0.64	0.58	0.71	0.65	0.63

ESIR method gets better. Moreover, although the asymptotic theory is based on the elliptical assumption of the covariates, the results of Model (B4) indicate that our method can be applied to a much wider range of distributions characterized by heavy tails.

To examine the influence of p, H and n on the estimation efficiency of the ESIR estimator, we consider the combinations of $n = 120, 200$, and 400, $p = 5, 10$, and 30, and $H = 5, 10, 20$, and 40 in Model (B1) for Cauchy distributed covariates. Simulation results are presented in Table 4 after 100 replicates.

In this setting, we avoid reporting the standard deviations and the averages of $R^2(\hat{\beta}_1)$ and $R^2(\hat{\beta}_2)$ to improve the clarity of the simulation results. From Table 4, we find that when n and H stay fixed, the larger p causes $R^2(\hat{\beta}_1)$ and $R^2(\hat{\beta}_2)$ to become smaller as high dimension reduces the estimation efficiency of both SIR and ESIR. However, our ESIR method seems to be not so sensitive to dimensionality as SIR. Looking at the rows of Table 4, $R^2(\hat{\beta}_1)$ and $R^2(\hat{\beta}_2)$ of the ESIR method decrease much slower when p gets larger. Secondly, when n gets larger, both SIR and ESIR tend to perform better, which fits our expectation. Lastly, it seems that the number of the slices does not have any significant impact on the estimation of both methods. It is not surprising, because Zhu and Ng (1995) and Zhu, Miao, and Peng (2006) also found such a phenomenon in their simulation studies for the SIR method. Table 5 reports similar results for t(2) distributed covariates with higher dimensions: $p = 50, 80, 100$. We do not present simulation results for other distributions or other models as they are quite similar to those of Tables 4 and 5.

6.3. ESIR With Robust Mean Estimators

In this section, the performances of ESIR with robust mean estimators are investigated. We replace the sample mean by the coordinate median, spatial median and median-of-means in Steps 1 and 3 of the algorithm, respectively. Compared with the original version, the new algorithm improves the performance

Table 5. Estimation of the central subspace for Model (B1) with $t(2)$ distributed covariates.

H		5			10			20			40		
p		50	80	100	50	80	100	50	80	100	50	80	100
$n = 120$													
SIR	$R^2(\hat{\beta}_1)$	0.32	0.20	0.14	0.31	0.16	0.11	0.21	0.10	0.05	0.11	0.06	0.02
	$R^2(\hat{\beta}_2)$	0.11	0.09	0.08	0.09	0.10	0.08	0.10	0.09	0.06	0.11	0.06	0.02
ESIR	$R^2(\hat{\beta}_1)$	0.51	0.43	0.41	0.46	0.40	0.41	0.46	0.39	0.36	0.44	0.33	0.42
	$R^2(\hat{\beta}_2)$	0.28	0.27	0.22	0.28	0.22	0.22	0.28	0.26	0.21	0.33	0.22	0.25
$n = 200$													
SIR	$R^2(\hat{\beta}_1)$	0.41	0.30	0.28	0.41	0.28	0.26	0.34	0.27	0.21	0.29	0.17	0.12
	$R^2(\hat{\beta}_2)$	0.10	0.07	0.06	0.12	0.09	0.07	0.11	0.08	0.08	0.08	0.08	0.06
ESIR	$R^2(\hat{\beta}_1)$	0.50	0.45	0.42	0.46	0.47	0.42	0.46	0.42	0.37	0.46	0.38	0.35
	$R^2(\hat{\beta}_2)$	0.27	0.24	0.24	0.25	0.30	0.26	0.33	0.30	0.25	0.37	0.29	0.24
$n = 400$													
SIR	$R^2(\hat{\beta}_1)$	0.50	0.40	0.36	0.48	0.40	0.37	0.48	0.39	0.36	0.44	0.37	0.31
	$R^2(\hat{\beta}_2)$	0.15	0.10	0.08	0.17	0.13	0.10	0.13	0.10	0.10	0.10	0.07	0.07
ESIR	$R^2(\hat{\beta}_1)$	0.52	0.47	0.46	0.59	0.53	0.48	0.62	0.51	0.39	0.60	0.44	0.38
	$R^2(\hat{\beta}_2)$	0.24	0.26	0.24	0.30	0.27	0.26	0.30	0.29	0.26	0.32	0.28	0.28

Table 6. Means and standard deviations of average R^2 for Model (A1).

	$p = 5$		$p = 10$		$p = 30$	
Method	Mean	STD	Mean	STD	Mean	STD
ESIR	0.54	0.34	0.47	0.34	0.49	0.36
CM-ESIR	0.89	0.18	0.80	0.23	0.64	0.29
SM-ESIR	0.89	0.18	0.82	0.21	0.66	0.32
CMM-ESIR	0.60	0.34	0.47	0.34	0.43	0.34
SMM-ESIR	0.66	0.31	0.54	0.34	0.43	0.34

Table 7. Means and standard deviations of average R^2 for Model (B1).

	$p = 5$		$p = 10$		$p = 30$	
Method	Mean	STD	Mean	STD	Mean	STD
ESIR	0.78	0.20	0.71	0.23	0.66	0.22
CM-ESIR	0.92	0.10	0.83	0.17	0.74	0.19
SM-ESIR	0.91	0.11	0.82	0.15	0.75	0.18
CMM-ESIR	0.85	0.15	0.76	0.18	0.66	0.23
SMM-ESIR	0.86	0.15	0.74	0.19	0.67	0.22

of ESIR, which will be shown clearly by the following simulation results.

Models (A1) and (B1) are employed to compare the performances of the new algorithm and the original one. For the new algorithm, coordinate median, spatial median, coordinate median-of-means and spatial median-of-means are exploited in the estimation. We name the above four approaches as CM-ESIR, SM-ESIR, CMM-ESIR and SMM-ESIR for simplicity. The predictor $\mathbf{X}$ in Models (A1) and (B1) is generated from the multivariate standard Cauchy distribution, and we set $n = 400$ and $p = 5, 10,$ and 30. The means and standard deviations of the average squared multiple correlation coefficient R^2 defined in Equation (6.1) are reported after 100 repetitions. Recall that a larger R^2 indicates more efficient estimation.

Tables 6 and 7 show that the methods with robust mean estimation outperform that with the traditional sample mean in most scenarios. Among the four new methods, CM-ESIR and SM-ESIR perform especially well with larger means of the average squared multiple correlation coefficient and small standard deviations. CMM-ESIR and SMM-ESIR do not perform so well, probably because the sample size is limited but we need to divide the sample for both the SIR and the means-of-median processes. Overall, using robust mean estimators in the ESIR algorithm improves the estimation efficiency.

6.4. Comparison With Existing Methods

In the literature, there exist a few other works that also relax the joint Gaussian distributions for the SIR approach which work for the low-dimensional setting. Among them, the contour projection proposed by Wang, Ni, and Tsay (2008) and Luo, wang, and Tsay (2009) and the weighted inverse regression estimation (WIRE) and sliced inverse median estimation (SIME) proposed by Dong, Yu, and Zhu (2015) enjoy a high level of popularity. Dong, Yu, and Zhu (2015) compared their proposed WIRE and SIME with the contour projection of SIR and demonstrated that their methods have better performances. Therefore, we compare ESIR with the two methods proposed by Dong, Yu, and Zhu (2015).

To conduct the comparison fairly, we copy the simulation setting used by Dong, Yu, and Zhu (2015). The following three models are considered:

$$\mathrm{I}: Y = 1 + 0.6X_1 - 0.4X_2 + 0.8X_3 + 0.2\varepsilon\,,$$
$$\mathrm{II}: Y = (1 + 0.1\varepsilon)X_1\,,$$
$$\mathrm{III}: Y = X_1/\{0.5 + (X_2 + 1.5)^2\} + 0.2\varepsilon\,,$$

where $\varepsilon \sim N(0, 1)$ is independent of the predictor $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$. Models I and II have structural dimension $K = 1$, and Model III has $K = 2$. Models I and III are homoscedastic while Model II is heteroscedastic. Notice that Model III is the same as Model (B1) in Section 6.2. Consider the following two mechanisms for generating $\mathbf{X}$:

(i) $\mathbf{X}$ is standard multivariate Cauchy;
(ii) $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$, where X_j is generated independently from a mixture of normal and uniform distributions: $X_j = 0.8N(0, 1) + 0.2Unif(-0.1, 0.1)$ for $j = 1, \ldots, p$.

The distribution of $\mathbf{X}$ is elliptically symmetric in (i) and non-elliptical in (ii). There will be outliers in Case (i) and inliers in Case (ii).

WIRE and SIME are compared with our proposed ESIR method. As done in Dong, Yu, and Zhu (2015), robust mean estimation is employed in all the three methods. The performance of each method is evaluated by the trace correlation employed by Dong, Yu, and Zhu (2015). Denote $\mathbf{B}$ as an orthonormal basis of the central subspace and $\widehat{\mathbf{B}}$ as the sample estimate of $\mathbf{B}$. The trace correlation (Ferré 1998) is defined as

$$r = \mathrm{trace}(P_{\mathbf{B}} P_{\widehat{\mathbf{B}}})/K\,,$$

where $P_{\mathbf{A}}$ denotes the projection matrix onto the column space of $\mathbf{A} \in \mathbb{R}^{p\times K}$. Larger values of trace correlation indicate better estimation.

We compare ESIR, WIRE, and SIME across Models I-III and two types of distributions of $\mathbf{X}$. Set $n = 200$ and $p = 5, 10, 30$. The means and standard deviations of the trace correlation r are calculated based on 200 repetitions.

Tables 8–10 report the averages and standard deviations of the trace correlation r with $p = 5$, $p = 10$ and $p = 30$, respectively. In Models I and II, ESIR outperforms WIRE and SIME in both generating schemes of the predictor $\mathbf{X}$ and in all $p = 5, 10, 30$ settings. Moreover, ESIR has a larger mean and a smaller standard deviation of trace correlation than WIRE and SIME in each case, and the advantage of ESIR becomes more obvious when the dimension increases. In Model III, WIRE and SIME perform a little better than ESIR in the $p = 5$ case, however the difference between these methods seems quite small. When the dimension goes higher, ESIR tends to outperform WIRE and SIME again. In summary, our proposed ESIR has better overall performances than existing robust SIR approaches.

7. Real Data Analysis

We exploit the Istanbul stock exchange dataset (*http://archive.ics.uci.edu/ml/datasets/ISTANBUL+STOCK+EXCHANGE*) to demonstrate the superiority of ESIR to SIR when the covariates are characterized with non-Gaussian and heavy-tailed features.

Table 8. Means and standard deviations of the trace correlation r with $p = 5$.

		Method					
		ESIR		WIRE		SIME	
Model	X	Mean	STD	Mean	STD	Mean	STD
I	(i)	**1.00**	0.00	1.00	0.01	0.99	0.03
	(ii)	**1.00**	0.00	1.00	0.00	1.00	0.00
II	(i)	**0.99**	0.01	0.95	0.03	0.93	0.05
	(ii)	**0.99**	0.00	0.99	0.01	0.99	0.01
III	(i)	0.90	0.08	0.91	0.07	**0.92**	0.06
	(ii)	0.94	0.04	**0.95**	0.03	0.95	0.04

Table 9. Means and standard deviations of the trace correlation r with $p = 10$.

		Method					
		ESIR		WIRE		SIME	
Model	X	Mean	STD	Mean	STD	Mean	STD
I	(i)	**1.00**	0.00	0.98	0.03	0.95	0.07
	(ii)	**1.00**	0.00	1.00	0.00	1.00	0.00
II	(i)	**0.99**	0.01	0.89	0.06	0.84	0.08
	(ii)	**0.99**	0.00	0.98	0.01	0.98	0.01
III	(i)	**0.80**	0.10	0.79	0.08	0.78	0.08
	(ii)	0.87	0.07	**0.88**	0.06	0.88	0.06

Table 10. Means and standard deviations of the trace correlation r with $p = 30$.

		Method					
		ESIR		WIRE		SIME	
Model	X	Mean	STD	Mean	STD	Mean	STD
I	(i)	**1.00**	0.00	0.79	0.10	0.69	0.11
	(ii)	**1.00**	0.00	0.98	0.04	0.97	0.05
II	(i)	**0.95**	0.01	0.68	0.08	0.61	0.10
	(ii)	**0.98**	0.01	0.88	0.05	0.87	0.06
III	(i)	**0.59**	0.09	0.48	0.09	0.49	0.08
	(ii)	**0.67**	0.08	0.63	0.08	0.63	0.08

Table 11. Normality tests.

Variable	Shapiro–Wilk	Kolmogorov–Smirnov
ISE	0.98***	0.05·
SP	0.94***	0.11***
DAX	0.97***	0.07*
FTSE	0.97***	0.07*
NIKKEI	0.98***	0.07*
BOVESPA	0.97***	0.06*
EU	0.97***	0.07*

There are 536 observations and 8 variables in the data set: Istanbul stock exchange national 100 index (ISE), Standard Poole 500 return index (SP), Stock market return index of Germany (DAX), Stock market return index of UK (FTSE), Stock market return index of Japan (NIKKEI), Stock market return index of Brazil (BOVESPA), MSCI European index (EU) and MSCI emerging markets index (EM). EM is chosen as the response variable and the other variables as the covariates to form a regression problem.

First, we explore the marginal distributions of the independent variables. Two normality tests, Shapiro–Wilk test and Kolmogorov–Smirnov test, are conducted to check the non-Gaussian feature of these variables. Table 11 summarizes the Shapiro–Wilk statistics and Kolmogorov–Smirnov statistics. We find that the covariates can all be considered as non-Gaussian distributed at the significance level of 0.05 except that the conclusion on the variable ISE is controversial. We then plot the empirical densities of the standardized covariates against the standard normal distribution to illustrate the heavy-tailed pattern. Figure 1 exhibits this heavy-tailed character pretty clearly. It can also be seen from this figure that all the covariates are symmetric about 0. Thus, we can readily apply the ESIR method to this dataset.

To determine the dimension K of the central subspace, the widely used marginal dimension test is applied. The number of slices is set to be 10. The test result suggests that $K = 2$ would be a proper choice. Therefore, we set the dimension of the central subspace to be $K = 2$ and the number of slices $H = 10$. After estimating the e.d.r. directions, we get two new indices: $\hat{\boldsymbol{\beta}}_1^{\mathrm{T}}\mathbf{X}$ and $\hat{\boldsymbol{\beta}}_2^{\mathrm{T}}\mathbf{X}$, then use them and $(\hat{\boldsymbol{\beta}}_1^{\mathrm{T}}\mathbf{X})^2$, $(\hat{\boldsymbol{\beta}}_2^{\mathrm{T}}\mathbf{X})^2$, and $(\hat{\boldsymbol{\beta}}_1^{\mathrm{T}}\mathbf{X}) \cdot (\hat{\boldsymbol{\beta}}_2^{\mathrm{T}}\mathbf{X})$ as explanatory variables to fit *EM*. The adjusted R-squared and F statistic are exploited to compare the performances of ESIR and SIR. We first use samples of the whole time period and then extend to investigate samples from three shorter periods which appear to have significant heavy tails (see Figure 2). The results are presented in Table 12. Obviously, our method outperforms SIR in all the four periods with significantly larger values of both R-squared and F statistic. This finding is complied with the simulation results above. It can be conjectured that the ESIR method would work better for returns of individual asset, futures and derivatives with higher risk.

8. Discussion

In this article, we propose the ESIR method for sufficient dimension reduction, which is a robust alternative to SIR for analyzing high-dimensional, elliptically distributed data. The main idea is to exploit the multivariate Kendall's tau matrix

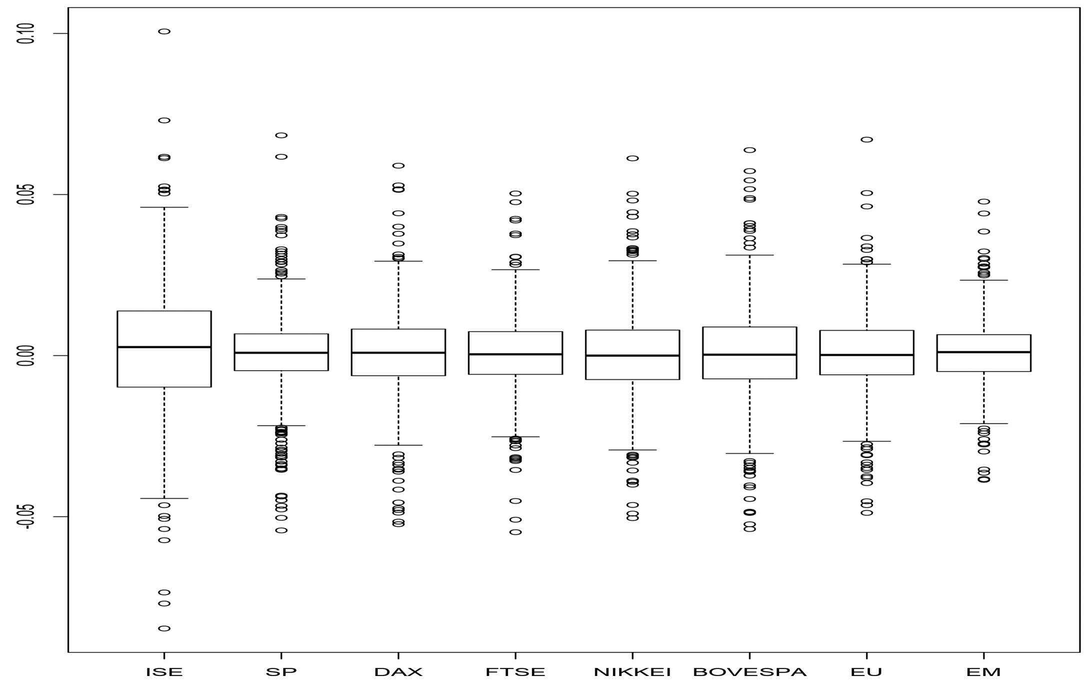

Figure 1. Boxplots of the stock returns.

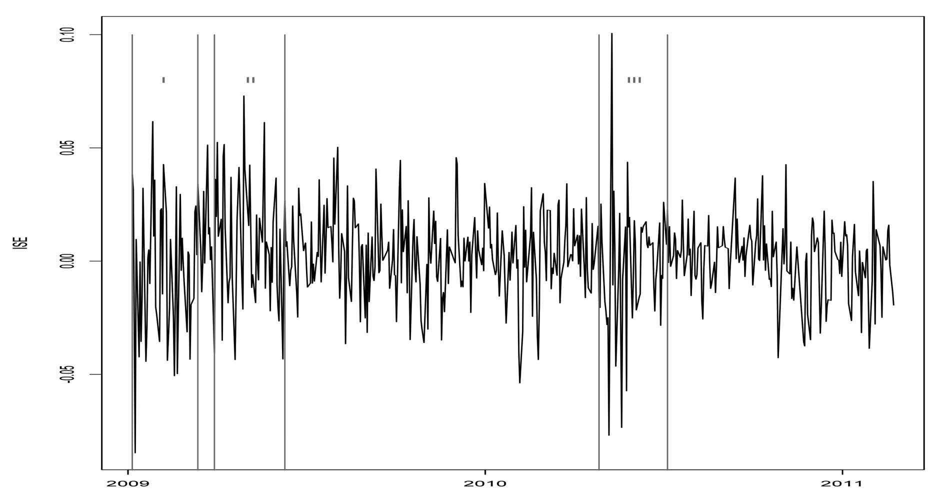

Figure 2. Three periods of the return of ISE.

Table 12. Regression results.

	Adjusted R^2		F statistic	
Time period	SIR	ESIR	SIR	ESIR
All	0.56	0.71	137.10	267.60
I	0.69	0.71	241.30	264.40
II	0.60	0.73	159.60	291.90
III	0.64	0.71	187.70	266.30
All	0.56	0.71	137.10	267.60

NOTE: The '***', '*', and '.' in cells represent the p-value less than 0.001, 0.05, and 0.15, respectively.

in a generalized eigenvalue problem to cope with the heavy-tailed elliptically distributed covariates. We then present a theorem to demonstrate the validity of the ESIR method for sufficient dimension reduction and give a simple and practicable algorithm for the ESIR method. The asymptotic behavior of the ESIR estimator is studied in the high-dimensional setting. Simulation results and real data studies demonstrate that ESIR significantly improves the estimation efficiency in the case of elliptically distributed covariates. Moreover, in the generalized eigenvalue framework, the proposed method can be easily extended to other sufficient dimension reduction methods such as SAVE, DR, and principal fitted components etc. Please refer to Li (2007) and Chen, Zou, and Cook (2010) for generalized eigenvalue problem. Lastly, our method is of vital importance for analyzing heavy-tailed financial, genomic and bioimaging data.

Appendix A: Proofs

A.1. Proof of Theorem 1

Proof. Following the conclusion of Theorem 3.1 in Li (1991), if we can prove that the linear space spanned by $\boldsymbol{\Sigma}\boldsymbol{\beta}_k (k = 1, \dots, K)$ is the same as the space spanned by $\mathbf{M}\boldsymbol{\beta}_k (k = 1, \dots, K)$, we are done.

For any vector $\boldsymbol{\gamma} \in \mathbb{R}^K$, let $\mathbf{B}^* = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_k)^{\mathrm{T}}$. Then the span of $\mathbf{M}\boldsymbol{\beta}_k (k = 1, \dots, K)$ can be written as $\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{B}^*\mathbf{M}$ and

$$\begin{aligned}
\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{B}^*\mathbf{M} &= \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{B}^* \cdot \left\{ \sum_{j=1}^{p} \lambda_j(\mathbf{M})\boldsymbol{\mu}_j(\mathbf{M})\boldsymbol{\mu}_j^{\mathrm{T}}(\mathbf{M}) \right\} \\
&= \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{B}^* \cdot \left\{ \sum_{j=1}^{p} \lambda_j(\mathbf{M})\boldsymbol{\mu}_j(\boldsymbol{\Sigma})\boldsymbol{\mu}_j^{\mathrm{T}}(\boldsymbol{\Sigma}) \right\} \\
&= \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{B}^* \cdot \left[\sum_{j=1}^{p} \mathbb{E}\left\{ \frac{\lambda_j(\boldsymbol{\Sigma})Q_j^2}{\lambda_1(\boldsymbol{\Sigma})Q_1^2 + \ldots + \lambda_p(\boldsymbol{\Sigma})Q_p^2} \right\} \boldsymbol{\mu}_j(\boldsymbol{\Sigma})\boldsymbol{\mu}_j^{\mathrm{T}}(\boldsymbol{\Sigma}) \right] \\
&:= s\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{B}^* \cdot \left\{ \sum_{j=1}^{p} \lambda_j(\boldsymbol{\Sigma})\boldsymbol{\mu}_j(\boldsymbol{\Sigma})\boldsymbol{\mu}_j^{\mathrm{T}}(\boldsymbol{\Sigma}) \right\} \\
&= (s\boldsymbol{\gamma})^{\mathrm{T}}\mathbf{B}^*\boldsymbol{\Sigma},
\end{aligned}$$

where the first equality comes from the spectral decomposition of $\mathbf{M}$, the second one is established by the property that $\boldsymbol{\mu}_j(\boldsymbol{\Sigma}) = \boldsymbol{\mu}_j(\mathbf{M})$ (see Marden 1999; Croux, Ollila, and Oja 2002; Oja 2010; Han and Liu 2016 for details), the third equality is given by Proposition 2.1 of Han and Liu (2016) and $\mathbf{Q} := (Q_1, \dots, Q_p)^{\mathrm{T}} \sim N_p(\mathbf{0}, \mathbf{I}_p)$. The last equality verifies our guess. □

A.2. Proof of Proposition 1

Proof. The proof for this proposition is mainly based on the results of Marden (1999). For any $\mathbf{X} \sim EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \xi)$, we have the decomposition below:

$$\mathbf{X} = \boldsymbol{\Omega}\mathbf{W} + \mathbf{b} \tag{A.1}$$

where $\boldsymbol{\Omega}$ is some orthogonal matrix, $\mathbf{W} \in \mathbb{R}^p$ is coordinatewise symmetric about 0, that is

$$\mathbf{G}\mathbf{W} \overset{d}{=} \mathbf{W} \tag{A.2}$$

for any matrix $\mathbf{G}$ with $G_{jj} \in \{1, -1\}$ and $G_{ij} = 0 (i \neq j)$ and $\mathbf{b}$ is some $p \times 1$ centering vector. We assume that $\mathrm{cov}(\mathbf{W})$ exists, and without loss of generality, that $\lambda_1 \geq \cdots \geq \lambda_p$ with $\lambda_i = \mathrm{var}(W_i)$ for $\mathbf{W} = (W_1, \dots, W_p)^{\mathrm{T}}$. Thus, we obtain $\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{X}) = \boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Omega}^{T}$, where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \dots, \lambda_p)$ with $\lambda_1 \geq \cdots \geq \lambda_p$.

For the vector $F(Y) = \mathbb{E}(\mathbf{X}|Y)$, from (A.1) we have

$$F(Y) = \mathbb{E}(\mathbf{X}|Y) = \mathbb{E}(\boldsymbol{\Omega}\mathbf{W} + \mathbf{b}|Y) = \boldsymbol{\Omega}\mathbb{E}(\mathbf{W}|Y) + \mathbf{b} := \boldsymbol{\Omega}F_{\mathbf{W}}(Y) + \mathbf{b}. \tag{A.3}$$

Then for $F_{\mathbf{W}}(Y) = \mathbb{E}(\mathbf{W}|Y)$, we can derive from Equation (A.2) that for any $\mathbf{G}$ with $G_{jj} \in \{1, -1\}$ and $G_{ij} = 0 (i \neq j)$,

$$\mathbf{G}F_{\mathbf{W}}(Y) = \mathbf{G}\mathbb{E}(\mathbf{W}|Y) = \mathbb{E}(\mathbf{G}\mathbf{W}|Y) = \mathbb{E}(\mathbf{W}|Y) = F_{\mathbf{W}}(Y). \tag{A.4}$$

Therefore, $F_{\mathbf{W}}(Y) = \mathbb{E}(\mathbf{W}|Y)$ is coordinatewise symmetric about 0.

By Proposition of Marden (1999) and Equations (A.3) and (A.4), we obtain

$$\mathbf{M}_{\mathbb{E}(\mathbf{X}|Y)} = \boldsymbol{\Omega}(p\boldsymbol{\Lambda}_1)\boldsymbol{\Omega}^{\mathrm{T}},$$

where $\boldsymbol{\Lambda}_1$ is a $p \times p$ diagonal matrix whose diagonal elements are the eigenvalues of $\mathbf{M}_{\mathbb{E}(\mathbf{X}|Y)}$. □

Supplementary Material

ESIR-supp: Additional proofs for Proposition 2 and Theorem 2. (pdf)
ESIR-code: R-code for the simulation. (R File)

Acknowledgments

The authors are grateful to the editor, the associate editor and two referees for their constructive comments and suggestions which have greatly improved our article.

Funding

Jia Zhang was supported in part by the National Natural Science Foundation of China (Grant Nos. 71991472 and 72003150), the China Postdoctoral Science Foundation, and the Joint Lab of Data Science and Business Intelligence at SWUFE. Xin Chen's research was supported in part by startup fund at the Southern University of Science and Technology and the National Natural Science Foundation of China (Grant No. 12071205). Wang Zhou's research was partially supported by a grant R-155-000-211-114 at the National University of Singapore.

References

Bura, E., and Cook, R. D. (2001a), "Extending Sliced Inverse Regression: The Weighted Chi-squared Test," *Journal of the American Statistical Association*, 96, 996–1003. [1204,1208]

——— (2001b), "Estimating the Structural Dimension of Regressions Via Parametric Inverse Regression," *Journal of the Royal Statistical Society*, Series B, 63, 393–410. [1204]

Bura, E., and Forzani, L. (2015), "Sufficient Reductions in Regressions With Elliptically Contoured Inverse Predictors," *Journal of the American Statistical Association*, 110, 420–434. [1206]

Bura, E., Duarte, S., and Forzani, L. (2016), "Sufficient Reductions in Regressions With Exponential Family Inverse Predictors," *Journal of the American Statistical Association*, 111, 1313–1329. [1206]

Chen, X., Cook, R. D., and Zou, C. L. (2015), "Diagnostic Studies in Sufficient Dimension Reduction," *Biometrika*, 102, 545–558. [1208,1209]

Chen, X., Zou, C. L., and Cook, R. D. (2010), "Coordinate-Independent Sparse Sufficient Dimension Reduction and Variable Selection," *The Annals of Statistics*, 38, 3696–3723. [1205,1207,1213]

Choi, K., and Marden, J. (1998), "A Multivariate Version of Kendall's τ," *Journal of Nonparametric Statistics*, 9, 261–293. [1205]

Cizek, P., Härdle, W. K., and Weron, R. (Eds.). (2005), *Statistical Tools for Finance and Insurance*. Springer Science and Business Media. Berlin: Springer. [1205]

Cook, R. D. (1994), "On the Interpretation of Regression Plots," *Journal of the American Statistical Association*, 89, 177–189. [1204]

——— (1996), "Graphics for Regressions with a Binary Response," *Journal of the American Statistical Association*, 91, 983–992. [1204]

——— (1998), *Regression Graphics: Ideas for Studying Regressions through Graphics*, New York: Wiley. [1204]

——— (2000), "SAVE: A Method for Dimension Reduction and Graphics in Regression," *Communications in Statistics, Part A, Theory and Methods*, 29, 2109–2121. [1204]

——— (2007), "Fisher Lecture: Dimension Deduction in Regression" (with discussion), *Statistical Science*, 22, 1–26. [1204]

Cook, R. D., and Forzani, L. (2009), "Likelihood-Based Sufficient Dimension Reduction," *Journal of the American Statistical Association*, 104, 197–208. [1204]

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. [1204]

Cook, R. D., and Weisberg, S. (1991), "Comment on 'Sliced Inverse Regression for Dimension Reduction' by K. C. Li," *Journal of the American Statistical Association*, 86, 328–332. [1204]

Croux, C., Ollila, E., and Oja, H. (2002),, "Sign and Rank Covariance Matrices: Statistical Properties and Application to Principal Components Analysis," in *Statistics for Industry and Technology*, 257–69. Birkhauser. [1206,1214]

Davis, C., and Kahan, W. M. (1970), "The Rotation of Eigenvectors by a Perturbation. III," *SIAM Journal on Numerical Analysis*, 7, 1–46. [1208]

Dong, Y., Yu, Z., and Zhu, L. (2015), "Robust Inverse Regression for Dimension Reduction," *Journal of Multivariate Analysis*, 134, 71–81. [1211]

Fan, J., Liu, H., and Wang, W. (2015), "Large Covariance Estimation Through Elliptical Factor Models," arXiv preprint arXiv:1507.08377. [1205]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space," *Journal of the Royal Statistical Society*, Series B, 70, 849-911. [1208]

Ferré, L. (1998), "Determining the Dimension in Sliced Inverse Regression and Related Methods," *Journal of the American Statistical Association*, 93, 132-140. [1211]

Ferré, L., and Yao, A. F. (2003), "Functional Sliced Inverse Regression Analysis," *Statistics*, 37, 475–488. [1204]

——— (2005), "Smoothed Functional Inverse Regression," *Statistica Sinica*, 15, 665–683. [1204]

Han, F., and Liu, H. (2018), "ECA: High-dimensional Elliptical Component Analysis in Non-Gaussian Distributions," *Journal of the American Statistical Association*, 113, 252-268. [1205,1206,1208,1209,1214]

Hsing, T., and Carroll, R. J. (1992), "An Asymptotic Theory for Sliced Inverse Regression," *The Annals of Statistics*, 20, 1040–1061. [1204,1207]

Hsing, T., and Ren, H. (2009), "An RKHS Formulation of the Inverse Regression Dimension-reduction Problem," *The Annals of Statistics*, 37, 726-755. [1204]

Hult, H., and Lindskog, F. (2002), "Multivariate Extremes, Aggregation and Dependence in Elliptical Distributions," *Advances in Applied Probability*, 34, 587–608. [1205]

Jing, B. Y., Shao, Q. M., and Wang, Q. (2003), "Self-Normalized Cramer-type Large Deviations for Independent Random Variables," *The Annals of Probability*, 31, 2167–2215.

Johnstone, I., and Lu, A. (2009), "On Consistency and Sparsity for Principal Components Analysis in High Dimensions," *Journal of the American Statistical Association*, 104, 682–693. [1205]

Li, B., and Wang, S. L. (2007), "On Directional Regression for Dimension Reduction," *Journal of the American Statistical Association*, 102, 997–1008. [1204]

Li, B., Zha, H., and Chiaromonte, F. (2005), "Contour Regression: A General Approach to Dimension Reduction," *The Annals of Statistics*, 33, 1580–1616. [1204]

Li, K. C. (1991), "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association*, 86, 316–342. [1204,1205,1206,1208,1209,1213]

——— (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. [1204]

Li, L. (2007), "Sparse Sufficient Dimension Reduction," *Biometrika*, 94, 603–613. [1205,1207,1213]

Li, R., Zhong, W., and Zhu, L. (2012), "Feature Screening Via Distance Correlation Learning," *Journal of the American Statistical Association*, 107, 1129–1139. [1208]

Li, Y., and Hsing, T. (2010), "Deciding the Dimension of Effective Dimension Reduction Space for Functional and High-dimensional Data," *The Annals of Statistics*, 38, 3028-3062. [1204]

Lin Q., Zhao Z., and Liu J. S. (accepted), "On Consistency and Sparsity of Sliced Inverse Regression in High Dimensions," *The Annals of Statistics*, arXiv:1507.03895. [1204,1205,1208]

Liu, L., Hawkins, D. M., Ghosh, S., and Young, S. S. (2003), "Robust Singular Value Decomposition Analysis of Microarray Data," *Proceedings of the National Academy of Sciences*, 100, 13167-13172. [1205]

Luo, R., Wang, H., and Tsai, C. L. (2009), "Contour Projected Dimension Reduction," *The Annals of Statistics*, 37, 3743-3778. [1211]

Marden, J. (1999), "Some Robust Estimates of Principal Components," *Statistics and Probability Letters*, 43, 349–359. [1206,1214]

Ma, Y., and Zhu, L. (2012), "A Semiparametric Approach to Dimension Reduction," *Journal of the American Statistical Association*, 107, 168–179. [1204]

Oja, H. (2010), *Multivariate Nonparametric Methods with R: An Approach based on Spatial Signs and Ranks*, vol. 199. New York: Springer. [1206,1214]

Pan, W., Wang, X., Xiao, W., and Zhu, H. (2018), "A Generic Sure Independence Screening Procedure," *Journal of the American Statistical Association*, DOI: 10.1080/01621459.2018.1462709 [1208]

Posekany, A., Felsenstein, K., and Sykacek, P. (2011), "Biological Assessment of Robust Noise Models in Microarray Data Analysis," *Bioinformatics*, 27, 807–814. [1205]

Rachev, S. T. (Ed.). (2003), *Handbook of Heavy Tailed Distributions in Finance: Handbooks in Finance*, vol. 1, Amsterdam: Elsevier. [1205]

Ruttimann, U. E., Unser, M., Rawlings, R. R., Rio, D., Ramsey, N. F., Mattay, V. S., ... and Weinberger, D. R. (1998), "Statistical Analysis of Functional MRI Data in the Wavelet Domain," *IEEE Transactions on Medical Imaging*, 17, 142-154. [1205]

Schott, J. R. (1994), "Determining the Dimensionality in Sliced Inverse Regression," *Journal of the American Statistical Association*, 89, 141-148. [1208]

Segal, M. R., Dahlquist, K. D. and Conklin, B. R. (2003), "Regression Approaches for Microarray Data Analysis,'' *Journal of Computational Biology*, 10, 961-980.

Sheng, W., and Yin, X. (2013), "Direction Estimation in Single-Index Models Via Distance Covariance," *Journal of Multivariate Analysis*, 122, 148–161. [1204]

——— (2016), "Sufficient Dimension Reduction Via Distance Covariance," *Journal of Computational and Graphical Statistics*, 25, 91–104. [1204]

Vershynin, R. (2010), "Introduction to the Non-Asymptotic Analysis of Random Matrices," arXiv preprint arXiv:1011.3027.

Vu, V. Q. and Lei, J. (2013), "Minimax Sparse Principal Subspace Estimation in High Dimensions," *The Annals of Statistics*, 41, 2905-2947. [1208]

Wang, H., Ni, L., and Tsai, C. L. (2008), "Improving Dimension Reduction Via Contour-Projection," *Statistica Sinica*, 299–311. [1211]

Wedin, P. A. (1972), "Perturbation Bounds in Connection with Singular Value Decomposition," *BIT Numerical Mathematics*, 12, 99–111. [1208]

Xia, Y., Tong, H., Li, W. K., and Zhu, L. (2002), "An Adaptive Estimation of Dimension Reduction Space" (with discussion), *Journal of the Royal Statistical Society*, Series B, 64, 363–410. [1204]

Yang, S. S. (1977), "General Distribution Theory of the Concomitants of Order Statistics," *The Annals of Statistics*, 5, 996–1002. [1207]

Yin, X., and Hilafu, H. (2015), "Sequential Sufficient Dimension Reduction for Large p, Small n Problems," *Journal of the Royal Statistical Society*, Series B, 77, 879-892.

Zhu, L. P., Li, L., Li, R., and Zhu, L. X. (2011), "Model-Free Feature Screening for Ultrahigh-dimensional Data," *Journal of the American Statistical Association*, 106, 1464-1475. [1208]

Zhu, L. X., and Fang, K. T. (1996), "Asymptotics for Kernel Estimate of Sliced Inverse Regression," *The Annals of Statistics*, 24, 1055–1068. [1204]

Zhu, L. X. and Ng, K. W. (1995), "Asymptotics of Sliced Inverse Regression," *Statistica Sinica*, 5, 727–736. [1204,1207,1210]

Zhu, L. X., Miao, B. Q., and Peng, H. (2006), "On Sliced Inverse Regression With High-Dimensional Covariates," *Journal of the American Statistical Association*, 101, 630–643. [1204,1205,1207,1208,1209,1210]

Zhu, L. X., Ohtaki, M., and Li, Y. X. (2006), "Extensions of Sliced Inverse Regression Based Algorithms," Technical Report, University of Hong Kong, Dept. of Statistics and Actuarial Science. [1204]