Journal of Nonparametric Statistics

ISSN: 1048-5252 (Print) 1029-0311 (Online) Journal homepage: https://www.tandfonline.com/loi/gnst20

Fourier transform approach for inverse dimension reduction method

Jiaying Weng & Xiangrong Yin

To cite this article: Jiaying Weng & Xiangrong Yin (2018) Fourier transform approach for inverse dimension reduction method, Journal of Nonparametric Statistics, 30:4, 1049-1071, DOI: 10.1080/10485252.2018.1515432

To link to this article: https://doi.org/10.1080/10485252.2018.1515432

JOURNAL OF NONPARAMETRIC STATISTICS
2018, VOL. 30, NO. 4, 1049–1071
https://doi.org/10.1080/10485252.2018.1515432

Taylor & Francis
Taylor & Francis Group

Fourier transform approach for inverse dimension reduction method

Jiaying Weng and Xiangrong Yin

Department of Statistics, University of Kentucky, Lexington, KY, USA

ABSTRACT

Estimating an inverse regression space is especially important in sufficient dimension reduction. However, it typically requires a tuning parameter, such as the number of slices in a slicing method or bandwidth selection in a kernel estimation approach. Such a requirement not only affects the accuracy of estimates in a finite sample, but also increases difficulties for multivariate models. In this paper, we use a Fourier transform approach to avoid such difficulties and incorporate multivariate models. We further develop a Fourier transform approach to deal with variable selection, categorical predictor variables, and large p, small n data. To test the dimension, asymptotic results are obtained. Simulation studies and data analysis show the efficacy of our proposed methods.

ARTICLE HISTORY
Received 11 August 2017
Accepted 19 August 2018

KEYWORDS
Central subspaces; Fourier transform; Inverse regression; Sufficient dimension reduction

1. Introduction

Sufficient dimension reduction (SDR) (Li 1991; Cook 1996) aims to find a few linear combinations of predictors so that using such linear combinations will preserve the regression information. In the regression problem, suppose that $Y \in \mathbb{R}$ is the response variable and $\mathbf{X} \in \mathbb{R}^p$ is the predictor vector. If $F(Y \mid \mathbf{X}) = F(Y \mid \boldsymbol{\eta}^{\mathrm{T}}\mathbf{X})$, where $\boldsymbol{\eta} \in \mathbb{R}^{p\times d}$, $d \leq p$, then the subspace spanned by the columns of $\boldsymbol{\eta}$ is called a dimension reduction subspace (DRS). We are interested in the central subspace (CS), $\mathcal{S}_{Y\mid\mathbf{X}}$, which is defined as the intersection of all DRSs if the intersection itself is a DRS. Under mild conditions (Cook 1996; Yin et al. 2008), CS exists and is unique. In this paper, we assume the existence of CS. Let $d = \dim(\mathcal{S}_{Y\mid\mathbf{X}})$ be the dimension of CS, and $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ be a basis of CS. Then $Y \perp\!\!\!\perp \mathbf{X}$ given $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$, where $\perp\!\!\!\perp$ indicates independence, is equivalent to saying the conditional distribution of Y given $\mathbf{X}$ is the same as the conditional distribution of Y given $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$. Along with this idea of CS, some specific subspaces focus on regression mean, variance and quantiles (Cook and Li 2002; Yin and Cook 2002; Zhu and Zhu 2009; Luo et al. 2014).

Many SDR methods have been developed over the past 30 years. Sliced inverse regression (SIR) (Li 1991) and sliced average variance estimation (SAVE) (Cook and Weisberg 1991) are the most well-known methods. SIR is preferred to recover a linear relationship between the response and predictors, while SAVE can handle a symmetric relationship

CONTACT Xiangrong Yin yinxiangrong@uky.edu Department of Statistics, University of Kentucky, 715 Rose Street, Lexington, KY 40536, USA

between them. Another approach is the principal Hessian direction (PHD) (Li 1992). SIR and SAVE employ the inverse regression of $\mathbf{X}$ given Y, while PHD is a correlation type of joint approach. All these methods use the spectral-decomposition-based procedure, which follows two steps. The first step is to construct a nonnegative definite symmetric dimension reduction matrix $M \in \mathbb{R}^{p\times p}$, called a kernel dimension reduction matrix. The second step involves conducting the spectral decomposition of a sample version $\hat{M}$ of M. Then, the first d eigenvectors corresponding to the first d largest eigenvalues of $\hat{M}$ are the estimator of the CS.

For SIR or SAVE, the number of slices has to be chosen, and the choice of this number could be problematic. Hsing and Carroll (1992) derived asymptotic properties for a special case where each slice had only two observations, which was generalised by Zhu and Ng (1995). The result of Zhu and Ng (1995) can be interpreted as the number of observations per slice must be large enough to yield efficient estimates, but still relatively small when compared with the sample size. This suggests that slicing schemes with too many slices that have too few observations per slice should be avoided. However, empirically it is hard to establish a useful rule for selecting the number of slices. To avoid such difficulties, Zhu et al. (2010b) developed the cumulative mean estimation, which uses a weighted average of SIR kernel matrices from all possible slicing schemes with two slices. Furthermore, Cook and Zhang (2014) proposed fused estimators by cumulating different number of slices: fused inverse regression estimator (FIRE) and degenerated inverse regression estimator (DIRE). Another improvement for SIR is to use Fourier transform (Zhu et al. 2010c). Fourier transform was first introduced by Zhu and Zeng (2006) to recover the dimensions in central mean subspace and CS.

The concept of SDR for multivariate response $\mathbf{Y} \in \mathbb{R}^q$ is simply to replace univatiate Y by $\mathbf{Y}$. The majority of SDR methods focuses on the univariate response, however, many methods have been developed for multivariate regression as well. [For instance, slicing the multi-dimensional $\mathbf{Y}$ into hypercubes similar to intervals in one-dimension, k-nearest neighbourhood mean approaches (Aragon 1997; Hsing 1999; Setodji and Cook 2004), and approaches combining all the marginal SDR for each component of $\mathbf{Y}$ to estimate the multivariate CS (Cook and Setodji 2003; Saracco 2005; Yin and Bura 2006).] Li et al. (2008) proposed a projective resampling (PR) method to avoid multivariate slicing while effectively estimating the CS. When data have categorical variables, but SDR is only on continuous predictors, then such an SDR approach leads to partial SDR (Chiaromonte et al. 2002; Li et al. 2003). SDR is quite useful for reducing predictors and helping to build a better model. However, it is still difficult to interpret the predictors in the model as the linear combination consists of all the original variables. To this end, SDR with penalisation can help to select important variables, leading to sufficient variable selection (SVS). One of the approaches is a general procedure by Li (2007), which developed a sparse SDR estimator for a general dimension reduction kernel matrix by transforming the eigenvalue-decomposition approach to a regression-type optimisation problem. Then a penalty term (such as a L^1 penalty) is added to shrink the number of parameters. Recently, Yin and Hilafu (2015) developed a sequential SDR and SVS procedure to deal with the large p, small n data with two effective algorithms, combining the techniques of SDR methods for the univariate response, multivariate responses, partial SDR and penalisation.

In this paper, we provide further developments for Fourier transform (FT) in inverse regression and focus on multivariate response. Differing from the forward motivation of

Zhu et al. (2010c), our approach gives more detailed illustration of their inverse regression link and significantly develops the idea. We have the following main contributions: (A). We provide a result regarding the choice of the number of FTs, only a finite number, much less than the sample size as suggested by Zhu et al. (2010c), which is sufficient enough. Indeed, empirically, 50 FTs are sufficient, and the results are quite stable. This will not only save computational time, but also ensure the accuracy of the estimate. (B). We obtain the asymptotic tests for determining dimensions for FT. (C). We develop a partial SDR for FT and obtain the respective asymptotic tests for estimating dimensions. (D). We further propose SVS in two useful cases: D1. For $n > p$, we use the idea of Li (2007) to develop a sparse SDR version of FT, which produces sparse and more accurate estimate. D2. Using the sequential SDR and SVS of Yin and Hilafu (2015), we develop a procedure of FT to deal with large p, small n data.

The article is organised as follows: Section 2 provides the theoretical justifications for FT estimator, comparison between SIR and FT in population and sample sense, properties of choice of the number of FTs, and algorithms for estimating CS, along with the test for dimension. Section 3 develops a method for estimating the partial SDR using FT approach. Section 4 proposes sufficient variable selection for two situations: large n, small p and large p, small n data. Section 5 presents simulation studies and a real data analysis. Section 6 summarises our conclusion. All proofs are included in the Appendix.

2. Methodology

2.1. Estimation method

This section introduces FT estimator. To facilitate our discussion, we use standardised predictor $\mathbf{Z}$ of $\mathbf{X}$, due to the equivalence of the CS of $\mathbf{Y} \mid \mathbf{X}$ and the CS of $\mathbf{Y} \mid \mathbf{Z}$ (Cook 1998). Let $\mathbf{Z} = \Sigma^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$, where $\boldsymbol{\mu}$ and Σ are the mean and covariance matrix of $\mathbf{X}$. Under the well-known linearity condition, $m(\mathbf{y}) = \mathrm{E}(\mathbf{Z} \mid \mathbf{Y} = \mathbf{y}) \in \mathcal{S}_{\mathbf{Y} \mid \mathbf{Z}}$ (Cook 1998). Thus, estimating the space spanned by $m(\mathbf{y})$ ($\mathcal{S}_{\mathrm{E}(\mathbf{Z} \mid \mathbf{Y})}$) is to recover part of the CS. Let $f_{\mathbf{Y}}(\mathbf{y})$ be the marginal density distribution of $\mathbf{Y}$. Then, FT of the density-weighted conditional mean $m(\mathbf{y})f_{\mathbf{Y}}(\mathbf{y})$ is $\psi(\boldsymbol{\omega}) = \int \mathrm{e}^{\mathrm{i}\boldsymbol{\omega}^{\mathrm{T}}\mathbf{y}} m(\mathbf{y}) f_{\mathbf{Y}}(\mathbf{y})\, \mathrm{d}\mathbf{y} = a(\boldsymbol{\omega}) + \mathrm{i}b(\boldsymbol{\omega}), \boldsymbol{\omega} \in \mathbb{R}^q$, where $a(\boldsymbol{\omega})$, $b(\boldsymbol{\omega})$ are the real, imaginary part of $\psi(\boldsymbol{\omega})$, respectively.

We claim that $\psi(\boldsymbol{\omega}) = \mathrm{E}(\mathrm{e}^{\mathrm{i}\boldsymbol{\omega}^{\mathrm{T}}\mathbf{Y}}\mathbf{Z})$ and $\mathcal{S}_{\mathrm{E}(\mathbf{Z} \mid \mathbf{Y})} = \mathrm{Span}\{\psi(\boldsymbol{\omega}), \boldsymbol{\omega} \in \mathbb{R}^q\}$. The first assertion is due to

$$\begin{aligned}
\psi(\boldsymbol{\omega}) &= \int \mathrm{e}^{\mathrm{i}\boldsymbol{\omega}^{\mathrm{T}}\mathbf{y}} m(\mathbf{y}) f_{\mathbf{Y}}(\mathbf{y})\, \mathrm{d}\mathbf{y} = \int \mathrm{e}^{\mathrm{i}\boldsymbol{\omega}^{\mathrm{T}}\mathbf{y}} \mathrm{E}(\mathbf{Z} \mid \mathbf{y}) f_{\mathbf{Y}}(\mathbf{y})\, \mathrm{d}\mathbf{y} \\
&= \int \mathrm{E}(\mathrm{e}^{\mathrm{i}\boldsymbol{\omega}^{\mathrm{T}}\mathbf{y}}\mathbf{Z} \mid \mathbf{y}) f_{\mathbf{Y}}(\mathbf{y})\, \mathrm{d}\mathbf{y} = \mathrm{E}[\mathrm{E}(\mathrm{e}^{\mathrm{i}\boldsymbol{\omega}^{\mathrm{T}}\mathbf{Y}}\mathbf{Z} \mid \mathbf{Y})] \\
&= \mathrm{E}(\mathrm{e}^{\mathrm{i}\boldsymbol{\omega}^{\mathrm{T}}\mathbf{Y}}\mathbf{Z}).
\end{aligned}$$

Note that $\mathcal{S}_{\mathrm{E}(\mathbf{Z} \mid \mathbf{Y})} = \mathrm{Span}\{m(\mathbf{y}), \mathbf{y} \in \mathrm{supp}(f_{\mathbf{Y}})\} = \mathrm{Span}\{m(\mathbf{y})f_{\mathbf{Y}}(\mathbf{y}), \mathbf{y} \in \mathrm{supp}(f_{\mathbf{Y}})\} \subseteq \mathcal{S}_{\mathbf{Y} \mid \mathbf{Z}}$, under the linearity condition. By its inverse transform of $\psi(\boldsymbol{\omega})$, then $m(\mathbf{y})f_{\mathbf{Y}}(\mathbf{y}) = (2\pi)^{-1} \int \mathrm{e}^{-\mathrm{i}\boldsymbol{\omega}^{\mathrm{T}}\mathbf{y}} \psi(\boldsymbol{\omega})\, \mathrm{d}\boldsymbol{\omega}$. Thus, $\mathcal{S}_{\mathrm{E}(\mathbf{Z} \mid \mathbf{Y})} = \mathrm{Span}\{\psi(\boldsymbol{\omega}), \boldsymbol{\omega} \in \mathbb{R}^q\} = \mathrm{Span}\{a(\boldsymbol{\omega}), b(\boldsymbol{\omega}), \boldsymbol{\omega} \in \mathbb{R}^q\}$, so the second assertion holds. Note that above derivation differs from the forward

illustration of Zhu and Zeng (2006), but does agree with their comment on inverse regression approach (right above Proposition 1, p. 1295, Zhu et al. 2010c). Although we give more details, both lead to the same estimator.

FT estimates the CS just as SIR does, but they might be different in estimation. In the population sense, SIR and FT estimate the space spanned by $\mathrm{E}(\mathbf{Z} \mid \mathbf{Y} = \mathbf{y})$, regardless of continuous or categorical $\mathbf{Y}$. That is,

$$\mathrm{Span}\{\psi(\boldsymbol{\omega}), \boldsymbol{\omega} \in \mathbb{R}^q\} = \mathrm{Span}\{\mathrm{E}(\mathbf{Z} \mid \mathbf{Y} = \mathbf{y}),\ \mathbf{y} \in \mathbb{R}^q\}.$$

When $\mathbf{Y}$ is a categorical variable, in the sample sense, SIR and FT are also equivalent (See the Appendix). That is, for categorical $\mathbf{Y}$, the left-hand side of the above equation does not gain any useful information by changing the number of $\boldsymbol{\omega}$, comparing with the right-hand side of the above equation. However, for continuous response $\mathbf{Y}$, empirical estimates for these two methods are different in accuracy, mainly due to the limited sample size. Note that the left-hand side (using FT) needs to choose the number of $\boldsymbol{\omega}$, while the right-hand side (using slices) needs to select the number of slices. The right-hand side has uncertainty for selecting the number of slices. Theoretically, it should choose a large number of slices due to its conditional mean, but practically it should use a small number of slices due to limited sample size. It is also well-known that the number of slices will greatly affect the accuracy of estimates. However, it seems that FT is quite stable for choosing the number of $\boldsymbol{\omega}$, as long as it is large enough.

2.2. Property of covering and choice of ω

In Section 2.1, we assume that $\boldsymbol{\omega}$s are given. Practically, we need the number and the value of $\boldsymbol{\omega}$ for estimation. Note that $\boldsymbol{\omega} \in \mathbb{R}^q$, but we cannot take the entire $\mathbb{R}^q$. Proposition 2.1 below, however, indicates that a finite number of $\boldsymbol{\omega} \in \mathbb{R}^q$ will be enough to recover the entire $\mathcal{S}_{\mathrm{E}(\mathbf{Z} \mid \mathbf{Y})}$. Yin and Li (2011) used a general dense class of functions of $\mathbf{Y}$ to estimate CS. FT is one of such dense classes, so the proof of Proposition 2.1 is similar to that of Theorem 2.2 (Yin and Li 2011). Hence, we omit its proof.

Proposition 2.1:

(1) *There exists a finite sequence of* $\boldsymbol{\omega}_j \in \mathbb{R}^q$, $j = 1, \ldots, t$, *such that* $\mathcal{S}_{\mathrm{E}(\mathbf{Z} \mid \mathbf{Y})} = \mathrm{Span}\{a(\boldsymbol{\omega}_1), b(\boldsymbol{\omega}_1), \ldots, a(\boldsymbol{\omega}_t), b(\boldsymbol{\omega}_t)\}$.
(2) *Consider a random sequence* $\boldsymbol{\omega}_j, j = 1, 2, \ldots$, *there exists an integer* t_0 *such that for all* $t \geq t_0$, $\mathcal{S}_{\mathrm{E}(\mathbf{Z} \mid \mathbf{Y})} = \mathrm{Span}\{a(\boldsymbol{\omega}_1), b(\boldsymbol{\omega}_1), \ldots, a(\boldsymbol{\omega}_t), b(\boldsymbol{\omega}_t)\}$.

Since Proposition 2.1 only states the result on $\mathcal{S}_{\mathrm{E}(\mathbf{Z} \mid \mathbf{Y})}$, we do not need the linearity condition. But it does if it is used for SDR. Part 1 of Proposition 2.1 indicates that the finite number of $\boldsymbol{\omega}$ is enough to recover $\mathcal{S}_{\mathrm{E}(\mathbf{Z} \mid \mathbf{Y})}$ and one could choose as small as half of the dimension of $\mathcal{S}_{\mathrm{E}(\mathbf{Z} \mid \mathbf{Y})}$. But typically, we do not know the dimension of $\mathcal{S}_{\mathrm{E}(\mathbf{Z} \mid \mathbf{Y})}$. Part 2 of Proposition 2.1 indicates that if the number of selected $\boldsymbol{\omega}$ is large enough, we can then recover $\mathcal{S}_{\mathrm{E}(\mathbf{Z} \mid \mathbf{Y})}$. This again in practice does not provide a useful rule. However, our simulations later show that when the number of $\boldsymbol{\omega}$ is large enough, the results are quite stable. Indeed, we find that 50 $\boldsymbol{\omega}$s is enough for capturing the structure of CS, as well as testing the dimension.

Another related issue is how to select $\boldsymbol{\omega}$. Zhu et al. (2010c) provide an argument to choose $\boldsymbol{\omega}$. For a multivariate $\mathbf{Y}$, we choose a small s, say $s = 0.1$, with $P(|\boldsymbol{\omega}^{\mathrm{T}}\mathbf{Y}| > \pi) \leq s$, then randomly generate $\boldsymbol{\omega} \sim N(\mathbf{0}, s\pi^2/\mathrm{E}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y})I)$. Our limited simulations indicate that such a method performed very stable.

2.3. Algorithm

In this section, we summarise what we have discussed above and show the algorithm for the estimate using sample. Let $\Psi = (a(\boldsymbol{\omega}_1), b(\boldsymbol{\omega}_1), \ldots, a(\boldsymbol{\omega}_t), b(\boldsymbol{\omega}_t))$, for some $t > 0$, and $V = \Psi\Psi^{\mathrm{T}}$ as the population kernel matrix. Recall, $\psi(\boldsymbol{\omega}) = a(\boldsymbol{\omega}) + ib(\boldsymbol{\omega})$. Let $(\mathbf{y}_i, \mathbf{x}_i)$, $i = 1, \ldots, n$ be a random sample, and assume that the dimension of $\mathcal{S}_{\mathrm{E}(\mathbf{Z}\,|\,\mathbf{Y})}$ is known as d. The algorithm of FT, similar to that of Zhu, Zhu and Wen (2010), is the following:

(1) Standardize $\mathbf{x}_i$: $\hat{\mathbf{z}}_i = \hat{\Sigma}_{\mathbf{X}}^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}})$, $i = 1, \ldots, n$, where $\bar{\mathbf{x}}$ is the sample mean, and $\hat{\Sigma}_{\mathbf{X}}$ is the sample covariance matrix of $\mathbf{X}$.
(2) Choose $\{\boldsymbol{\omega}_j\}_{j=1}^t$ as in Section 2.2 and for each $\boldsymbol{\omega}_j$, calculate sample version of $\psi(\boldsymbol{\omega}_j)$:

$$\hat{\psi}(\boldsymbol{\omega}_j) = \frac{1}{n}\sum_{k=1}^{n} \mathrm{e}^{\mathrm{i}\boldsymbol{\omega}_j^{\mathrm{T}}\mathbf{y}_k}\hat{\mathbf{z}}_k,$$

and $\hat{a}(\boldsymbol{\omega}_j) = \mathrm{Real}(\hat{\psi}(\boldsymbol{\omega}_j))$ and $\hat{b}(\boldsymbol{\omega}_j) = \mathrm{Image}(\hat{\psi}(\boldsymbol{\omega}_j))$.
(3) Form $\hat{\Psi}$ and $\hat{V}$ as

$$\hat{\Psi} = \{\hat{a}(\boldsymbol{\omega}_j), \hat{b}(\boldsymbol{\omega}_j)\}_{j=1}^t, \quad \hat{V} = \hat{\Psi}\hat{\Psi}^{\mathrm{T}},$$

where $\hat{\Psi}$ is a $p \times 2t$ matrix and $\hat{V}$ is a $p \times p$ sample kernel matrix.
(4) The first d eigenvectors $(\hat{\eta}_i,\ i = 1, \ldots, d)$ of $\hat{V}$ corresponding to the first d largest eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_d$ are the estimated directions of $\mathcal{S}_{\mathrm{E}(\mathbf{Z}\,|\,\mathbf{Y})}$. Transform back to the $\mathbf{X}$ scale, $\hat{\beta}_i = \hat{\Sigma}_{\mathbf{X}}^{-1/2}\hat{\eta}_i$, $i = 1, \ldots, d$.

2.4. Testing methods for dimension

Previously we assume that d is known. However, practically, we do need to estimate d. We then construct the following statistic

$$\hat{\Lambda}_m = n\sum_{j=m+1}^{p} \hat{\lambda}_j$$

to test the hypothesis of the form $d = m$ versus $d > m$. The value of m begins with 0, so we test $d = m$ by comparing sample $\hat{\Lambda}_m$ with the quantile of the asymptotic distribution of $\hat{\Lambda}_m$ under the null hypothesis $d = m$. If we fail to reject, then $d = m$, otherwise we increase m by 1 and continue the same process until we fail to reject. The asymptotic distribution of $\hat{\Lambda}_d$ is stated the below Proposition 2.2, of which proof is in the Appendix. Again, Proposition 2.2 is stated in terms of $\mathcal{S}_{\mathrm{E}(\mathbf{Z}\,|\,\mathbf{Y})}$, the linearity condition is not necessary. However, it does need this condition for its use of SDR.

Proposition 2.2: *Let $d = \dim[\mathcal{S}_{\mathrm{E}(\mathbf{Z}\,|\,\mathbf{Y})}]$ and assume that $2t > d+1$ and $p > d$. Then the asymptotic distribution of $\hat{\Lambda}_d$ is the same as the distribution of*

$$C = \sum_{i=1}^{(p-d)(2t-d)} \lambda_i C_i,$$

where the C_is are independent chi-square random variables each with one degree of freedom and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{(p-d)(2t-d)}$ are eigenvalues of the covariance matrix Ω, where Ω is defined in the Appendix.

One can directly obtain the distribution of the weighted Chi-square Statistic C, however, simplification is possible. Following Bentler and Xie (2000), we consider two types of simplified test statistics.

Scaled Statistic: $\bar{T}_m = [\mathrm{trace}(\hat{\Omega}_n)/p^*]^{-1} n \sum_{j=m+1}^{p} \hat{\lambda}_j \sim \chi^2_{p^*}$, where $\hat{\Omega}_n$ is a consistent estimator of Ω and $p^* = (p-m)(2t-m)$.

Adjusted Statistic: $\tilde{T}_m = [\mathrm{trace}(\hat{\Omega}_n)/d^*]^{-1} n \sum_{j=m+1}^{p} \hat{\lambda}_j \sim \chi^2_{d^*}$, where $d^* = [\mathrm{trace}(\hat{\Omega}_n)]^2/\mathrm{trace}(\hat{\Omega}_n^2)$.

Sparse Eigen-Decomposition estimation (SED) (Zhu et al. 2010a) is another method to estimate d. We sketch SED here. Let $\hat{V}$ be the sample kernel matrix in Section 2.3. The SED procedure is the following:

$$(\hat{\boldsymbol{\lambda}}, \hat{\alpha}, \hat{\beta}) = \arg\min_{\lambda,\alpha,\beta} n \left\| \hat{V} - \sum_{i=1}^{p} \lambda_i \boldsymbol{\alpha}_i \boldsymbol{\beta}_i^{\mathrm{T}} \right\|^2 + l_n \sum_{i=1}^{p} \hat{w}_i |\lambda_i|,$$

subject to $\beta^{\mathrm{T}}\beta = I_p$, where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^{\mathrm{T}}$ be a $p \times 1$ vector, $\alpha = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p)$ and $\beta = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)$ be $p \times p$ matrices, $\hat{\boldsymbol{w}} = (\hat{w}_1, \ldots \hat{w}_p)^{\mathrm{T}}$ be a known weight vector. The tuning parameter, l_n, is selected by typical AIC and BIC as suggested by Zhu et al. (2010a). The number of dimensions is equal to the number of nonzero values of $\hat{\boldsymbol{\lambda}}$.

3. Partial central subspace

When predictors consist of both continuous and categorical variables, but SDR focuses on the continuous predictors, it then leads to partial SDR (Chiaromonte et al. 2002). In this section, we extend FT to partial SDR. Without loss of generality, let W be the categorical variable with K levels. Chiaromonte et al. (2002) defined the partial CS to be the intersection of all subspaces spanned by $\boldsymbol{\eta} \in \mathbb{R}^{p\times d}$ such that $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \,|\, (\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}, W)$, if the intersection itself also satisfies such a condition. Let $\mathcal{S}^W_{\mathbf{Y}\,|\,\mathbf{X}}$ be the partial CS, then $\mathcal{S}^W_{\mathbf{Y}\,|\,\mathbf{X}} = \bigoplus_{k=1}^{K} \mathcal{S}_{\mathbf{Y}_k\,|\,\mathbf{X}_k}$, where $\mathcal{S}_{\mathbf{Y}_k\,|\,\mathbf{X}_k}$ is the CS conditioning on level k.

Suppose that for each group, the mean and covariance matrix of $\mathbf{X}_k$ are $\boldsymbol{\mu}_k$ and Σ_k. To facilitate the discussion, we further assume that the covariance structures are the same across each level, that is, $\Sigma_k = \Sigma_{pool},\ k = 1, \ldots, K$. Let $\mathbf{Z}_k = \Sigma_{pool}^{-1/2}(\mathbf{X}_k - \boldsymbol{\mu}_k)$, then $\mathcal{S}^W_{\mathbf{Y}\,|\,\mathbf{X}} = \Sigma_{pool}^{-1/2} \bigoplus_{k=1}^{K} \mathcal{S}_{\mathbf{Y}_k\,|\,\mathbf{Z}_k}$. For each level, we construct the kernel matrix V_k and combine them into an overall kernel matrix: the partial kernel matrix $V^W = \sum_{k=1}^{K} P(W = k) V_k$. Suppose that the linearity and coverage conditions for each level hold:

- Linearity: $\mathrm{E}(\mathbf{Z}_k \mid P_{\mathcal{S}_{\mathbf{Y}_k|\mathbf{Z}_k}}\mathbf{Z}_k) = P_{\mathcal{S}_{\mathbf{Y}_k|\mathbf{Z}_k}}\mathbf{Z}_k$, for $k = 1, \ldots, K$.
- Coverage: $\mathrm{Span}(V^W) = \bigoplus_{k=1}^K \mathrm{Span}(V_k) = \bigoplus_{k=1}^K \mathcal{S}_{\mathbf{Y}_k|\mathbf{Z}_k}$.

Assume that the dimension of $\mathcal{S}^W_{\mathbf{Y}|\mathbf{X}}$, d, is known, we have the following algorithm for estimating $\mathcal{S}^W_{\mathbf{Y}|\mathbf{X}}$. The algorithm is similar to Chiaromonte et al. (2002) except applying our new partial kernel matrix. The estimate from the following steps is referred as the partial Fourier transform (PFT).

(1) For each level k, $\bar{\boldsymbol{x}}_k$ and $\hat{\Sigma}_k$ are the sample mean and covariance matrix of $\mathbf{X}_k$, the common covariance matrix is $\hat{\Sigma}_{pool} = \sum_{k=1}^K (n_k/n)\hat{\Sigma}_k$ and $\hat{\mathbf{z}}_{ik} = \hat{\Sigma}_{pool}^{-1/2}(\mathbf{x}_{ik} - \bar{\mathbf{x}}_k)$, $i = 1, \ldots, n_k$ and $k = 1, \ldots, K$.
(2) Apply the algorithm in Section 2.3 to obtain the sample kernel matrix for each level k: $\hat{V}_k$, and then $\hat{V}^W = \sum_{k=1}^K (n_k/n)\hat{V}_k$.
(3) The first d eigenvectors $(\hat{\eta}_i,\ i = 1, \ldots, d)$ of $\hat{V}^W$ corresponding to the first d largest eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_d$ are the estimates. Transform back to the $\mathbf{X}$ scale, $\hat{\beta}_i = \hat{\Sigma}_{pool}^{-1/2}\hat{\eta}_i,\ i = 1, \ldots, d$.

To estimate d of PFT, we construct a test statistic

$$\hat{\Lambda}^W_m = n \sum_{j=m+1}^{p} \hat{\lambda}_j.$$

Proposition 3.1: *Under the linearity and coverage conditions for partial SDR, let* $d = \dim[\mathcal{S}^W_{\mathbf{Y}|\mathbf{X}}]$ *and assume that* $2\sum t_k > Kd + 1$ *and* $p > d$. *Then the asymptotic distribution of* $\hat{\Lambda}^W_d$ *is the same as the distribution of*

$$C = \sum_{i=1}^{(p-d)(2\sum t_k - Kd)} \lambda_i C_i$$

where the C_i*s are independent chi-square random variables each with one degree of freedom and* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{(p-d)(2\sum t_k - Kd)}$ *are eigenvalues of the covariance matrix* Ω^W, *where* Ω^W *is defined in the appendix.*

4. Sufficient variable selection

In many cases, only a few predictors contribute for the model, which leads to sparse model. SDR with penalisation is helpful to choose such variables. In this section, we extend FT for sufficient variable selection via the penalised approach. We consider two different cases: the traditional large n, small p data, and the modern large p, small n data.

Large n, Small p: We adopt a general sparse SDR via penalty approach developed by Li (2007): $\tilde{V}\tilde{\eta}_i = \rho_i \Sigma \tilde{\eta}_i$, for $i = 1, \ldots, p$, where $\tilde{V} = \Sigma^{1/2} V \Sigma^{1/2}$ is a symmetric kernel SDR matrix; Σ is the covariance matrix; vector $\tilde{\eta}_1, \ldots, \tilde{\eta}_p$ are eigenvectors satisfying $\tilde{\eta}_i^{\mathrm{T}} \Sigma \tilde{\eta}_j = 1$ if $i = j$, and 0 if $i \neq j$; and $\rho_1 \geq \cdots \geq \rho_p \geq 0$ are corresponding eigenvalues. Then the eigenvalue-decomposition approach via penalty term becomes an optimisation

problem for sparse SDR as follows:

$$\min_{\alpha,\beta}\left(\sum_{i=1}^{p}\|\Sigma^{-1}v_i-\alpha\beta^{\mathrm{T}}v_i\|_{\Sigma}^2+\lambda_2\mathrm{trace}(\beta^{\mathrm{T}}\Sigma\beta)+\sum_{j=1}^{d}\lambda_{1j}|\beta_j|_1\right),$$

subject to $\alpha^{\mathrm{T}}\Sigma\alpha=I_d$, where $\lambda_{1j}\geq 0, j=1,\ldots,d$ are the tuning parameters, and v_i, $i=1,\ldots,p$ are the columns of $\tilde{V}^{1/2}$.

The algorithm of Li (2007) can be summarised as below:

(1) Initialize α and β using the sample kernel matrix in Section 2.3.
(2) Given α, update β as below:

$$\hat{\beta}_{\alpha j}=\arg\min_{\beta_j}(\|y^*-x^*\beta_j\|^2+\lambda_{1j}|\beta_j|_1),$$

where $x^*=\begin{bmatrix}\tilde{V}^{1/2}\\ \sqrt{\lambda_2}\Sigma^{1/2}\end{bmatrix}_{2p\times p}$, $y^*=\begin{bmatrix}\tilde{V}^{1/2}\alpha_j\\ 0\end{bmatrix}_{2p\times 1}$.
(3) Given β, let U_α, D_α, and V_α denote the matrices from the singular value decomposition of the matrix $\Sigma^{-1/2}\tilde{V}\beta$, then $\hat{\alpha}=\Sigma^{-1/2}U_\alpha V_\alpha^{\mathrm{T}}$.
(4) Continue steps 2 and 3 until β converges.

Typically, we need to fix λ_{1j} and λ_2 in the above algorithm. The final selection of tuning parameters of λ_{1j} and λ_2 can be determined by AIC and BIC (Li 2007). For our purpose, we simply use FT kernel matrix to replace $\tilde{V}$, and denote such a procedure as S-FT.

Large p, Small n: Yin and Hilafu (2015) proposed a sequential SDR (SSDR) for such a problem. We extend FT in their algorithm. Note that the algorithm of Yin and Hilafu (2015) is based on the following result.

Proposition 4.1 (Yin and Hilafu 2015): *If* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *are random vectors,* $B^{\mathrm{T}}\mathbf{X}_1$ *is a linear combination of* $\mathbf{X}_1$*, where B is a matrix, then either* (*a*) *or* (*b*) *implies* (*c*) *below:*

(a) $\mathbf{X}_1 \perp\!\!\!\perp (\mathbf{X}_2,\mathbf{Y}) \mid B^{\mathrm{T}}\mathbf{X}_1$;
(b) $\mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2 \mid \{B^{\mathrm{T}}\mathbf{X}_1,\mathbf{Y}\}$ *and* $\mathbf{X}_1 \perp\!\!\!\perp \mathbf{Y} \mid B^{\mathrm{T}}\mathbf{X}_1$;
(c) $\mathbf{X}_1 \perp\!\!\!\perp \mathbf{Y} \mid \{B^{\mathrm{T}}\mathbf{X}_1,\mathbf{X}_2\}$.

Statement (c) is very important, if it is true, then $p(\mathbf{Y}\mid\mathbf{X}_1,\mathbf{X}_2)=p(\mathbf{Y}\mid\mathbf{X}_1,\mathbf{X}_2,B^{\mathrm{T}}\mathbf{X}_1)=p(\mathbf{Y}\mid\mathbf{X}_2,B^{\mathrm{T}}\mathbf{X}_1)$. Thus, if the dimension of $B^{\mathrm{T}}\mathbf{X}_1$ is less than $\mathbf{X}_1$, we achieved dimension reduction without loss of any information. To force statement (c), we may use statement (a) or statement (b). Write $\mathbf{X}=(\mathbf{X}_1,\mathbf{X}_2)$, and choose $\mathbf{X}_1$ with dimension $p_1<n$. Then reduce $\mathbf{X}_1$ to $B^{\mathrm{T}}\mathbf{X}_1$, and replace $\mathbf{X}$ with $(B^{\mathrm{T}}\mathbf{X}_1,\mathbf{X}_2)$ as new $\mathbf{X}$. Keep doing this until there is no more reduction. To find $B^{\mathrm{T}}\mathbf{X}_1$, Path I procedure (Yin and Hilafu 2015) uses statement (a) when the response variable is continuous. This procedure needs to construct $B^{\mathrm{T}}\mathbf{X}_1$ using regression $(\mathbf{X}_2,\mathbf{Y})$ on $\mathbf{X}_1$. On the other hand, when dealing with the categorical response, statement (b) is the choice which is called Path II procedure by Yin and Hilafu (2015). Path II conducts the partial SDR for regression $\mathbf{X}_2$ on $\mathbf{X}_1$ given $\mathbf{Y}$, and the usual SDR of $\mathbf{Y}$ on $\mathbf{X}_1$. Because of the categorical response, FT is equivalent to SIR, we only use Path I to construct an estimate. For clarity, we illustrate the algorithm of Path I of Yin and Hilafu for FT below.

(1) Order the predictors using the distance correlation in Li et al. (2012).
(2) Decompose $\mathbf{X} \in \mathbb{R}^p$ into $\mathbf{X}^{\mathrm{T}} = (\mathbf{X}_1^{\mathrm{T}}, \mathbf{X}_2^{\mathrm{T}})$, where $\mathbf{X}_1$ is a $p_1 \times 1$ vector such that $n > p_1$, and consider the problem of $\mathbf{X}_1 \perp (\mathbf{X}_2, Y) \mid \beta_1^{\mathrm{T}}\mathbf{X}_1$.
(3) For SDR step, apply the method in Section 2.3 to new response $\mathbf{Y}_{new}^{\mathrm{T}} = (\mathbf{X}_2^{\mathrm{T}}, \mathbf{Y})$ given $\mathbf{X}_1$, and find the reduced variable $\beta_1^{\mathrm{T}}\mathbf{X}_1$; For SVS step, apply multivariate regression with penalisation to the problem of $\mathbf{Y}_{new}^{\mathrm{T}} \mid \mathbf{X}_1$, and find the reduced variable $\beta_1^{\mathrm{T}}\mathbf{X_1}$.
(4) Replace predictors $\mathbf{X}$ by $(\beta_1^{\mathrm{T}}\mathbf{X}_1, \mathbf{X}_2)$ and go back to step 1, until there is no further reduction.

We will compare the original SSDR using SIR (SSDR-SIR) and SSDR using FT (SSDR-FT) in the simulation for Path I.

5. Numerical study

Suppose that $\hat{B} = (\hat{\beta}_1, \ldots, \hat{\beta}_d)$ is the estimate of a $p \times d$ matrix B, and both $\hat{B}$ and B are orthogonal matrices. We use following criteria to measure the accuracy of the estimates.

(1) Let ρ_i^2's be the eigenvalues of matrix $\hat{B}^{\mathrm{T}}BB^{\mathrm{T}}\hat{B}$ for $i = 1, \ldots, d$: the vector correlation coefficient is $r_1 = \sqrt{|\hat{B}^{\mathrm{T}}BB^{\mathrm{T}}\hat{B}|} = |\prod_{i=1}^{d} \rho_i|$ and the trace correlation is $r_2 = \sqrt{\sum_{i=1}^{d} \rho_i^2/d}$ (Ye and Weiss 2003). The bigger the r_1 or r_2, the better the estimate.
(2) Define $\Delta(B, \hat{B}) = \|\hat{B}\hat{B}^{\mathrm{T}} - BB^{\mathrm{T}}\|$ (Li et al. 2005). We use two ways to calculate $\|.\|$: (a) $\Delta_m(A) = \|A\|$ is the maximum singular value of A, and (b) $\Delta_f(A) = \|A\|$ is the Frobenius norm as $\Delta_f(A) = \sqrt{\mathrm{trace}(AA^{\mathrm{T}})}$. The smaller the $\Delta_m(A)$ or $\Delta_f(A)$, the better the estimate.

For SVS, we use true positive rate (TPR) and false positive rate (FPR): TPR is the number of correctly identified active predictors to the number of truly active predictors, and FPR is the number of falsely identified active predictors to the total number of inactive predictors to compare different methods. Better estimates have bigger TPRs and smaller FPRs.

5.1. Simulations

In this section, we illustrate the advantages of FT with six models. Each model has a different purpose. We use Model 5.1 to assess if the number of $\boldsymbol{\omega}$'s in FT could affect estimate accuracy and Model 5.2 to compare FT with SIR, IRE (Cook and Ni 2005), FIRE and DIRE and, further to compare S-FT with S-SIR. We use Model 5.3 to estimate the dimension using the Weighted Chi-square, Scaled, Adjusted Statistic and SED and Model 5.4 for multivariate regression. We use Model 5.5 to compare partial SDR using SIR (PSIR) (Chiaromonte et al. 2002) and PFT. Finally, Model 5.6 is used for a large p, small n problem.

Model 5.1: *$Y = X_1 + 0.5X_2^2$, with $p = 5$, $n = 800$ and $d = 2$. Predictors $X_1, X_3, X_5 \overset{\text{iid}}{\sim} N(0,1)$, and $X_2 = X_1 + Z$ where $Z \sim N(0,1)$ and $X_4 = (1 + X_2)Z$. Let $\{e_i\}$ be $p \times 1$ vectors whose ith entry is 1 and other entries are 0. Then $B = (e_1, e_2)$.*

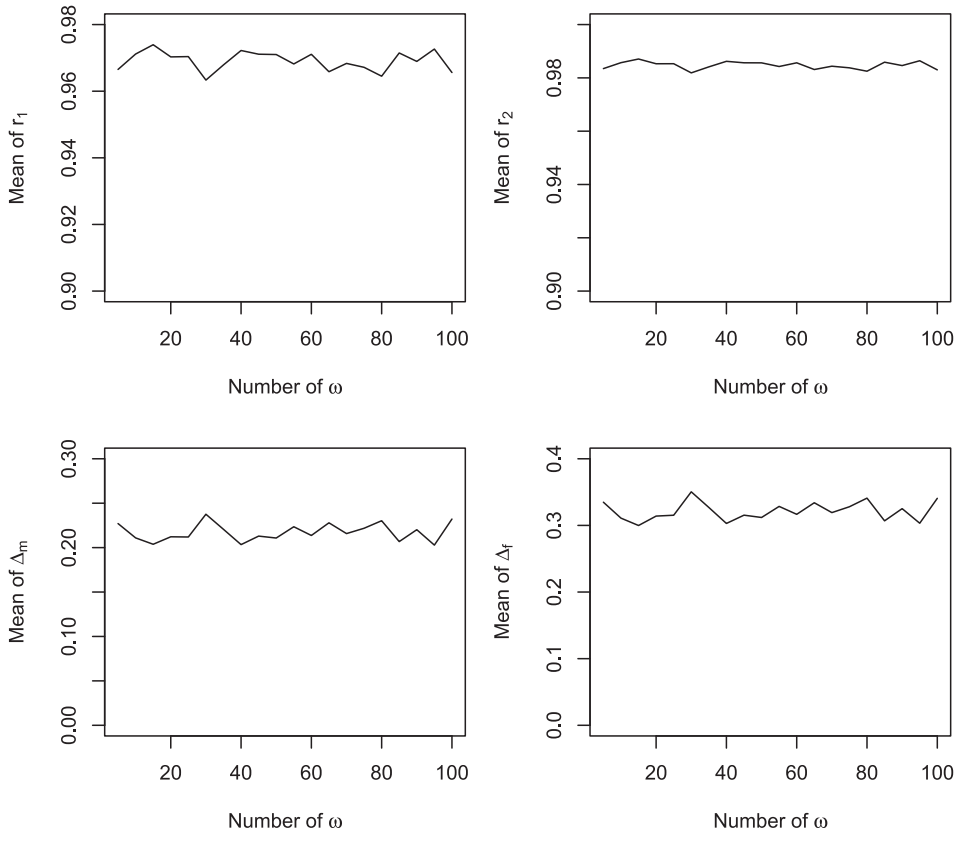

Figure 1. Mean values of respective r_1, r_2, Δ_m and Δ_f over 100 simulated data vs sizes of $\boldsymbol{\omega}$: $\{5, \ldots, 100\}$ in Model 1.

Figure 1 plots mean values of respective r_1, r_2, Δ_m and Δ_f over 100 simulated data vs sizes of $\boldsymbol{\omega}$: $\{5, 10, 15, \ldots, 100\}$. It shows that FT has high accuracy, and its estimates keep the same magnitude for the different number of $\boldsymbol{\omega}$'s. This seems consistent with the result of Proposition 2.1. Hence, as long as the size of $\boldsymbol{\omega}$ is large enough, estimates of the CS are accurate and stable.

Model 5.2: *This is the first example of* Cook and Zhang (2014). $Y = |\sin X_1| + 0.2\epsilon$, *with* $d = 1$ *and* $B = e_1$. *Predictors* $\mathbf{X}_i \sim \frac{1}{4}N_p(\boldsymbol{\mu}_1, \Sigma_1) + \frac{1}{2}N_p(\boldsymbol{\mu}_2, \Sigma_2) + \frac{1}{4}N_p(\boldsymbol{\mu}_3, \Sigma_3)$, *where* $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_3 = (1, 0, \ldots, 0)^{\mathrm{T}}$, $\boldsymbol{\mu}_2 = (2, 0, \ldots, 0)^{\mathrm{T}}$, $\Sigma_1 = \Sigma_2 = \sqrt{0.1}I_p$ *and* $\Sigma_3 = \sqrt{10}I_p$. *Let* $p = 15$, $n = 400$, *and* ϵ *is a uniform* (0,1).

We compare SIR, IRE, FIRE, DIRE, and FT for this model. The number of slices for SIR and IRE are $\{3, 4, \ldots, 15\}$. For FIRE and DIRE, we fuse $H = \{3, 4, \ldots, 15\}$, while for FT, the size of $\boldsymbol{\omega}$ is 50. Figure 2 plots mean values of respective r_1, r_2, Δ_m and Δ_f over 100 simulated data vs the number of slices from 3 to 15. We see that the results of SIR and IRE change with different slices, indicating that the choice of number of slices is important. FIRE and DIRE combine different slices together, thus they are constant lines. Regardless, FT has the largest values of r_1 and r_2, and the smallest Δ_m and Δ_f compared with the other

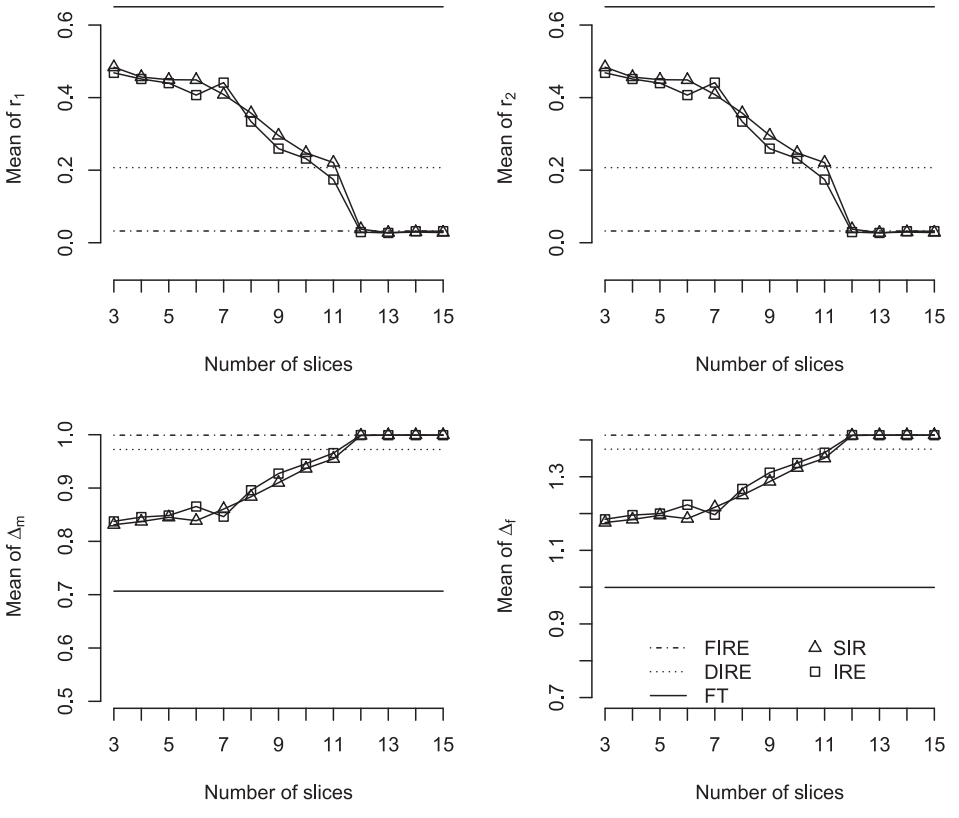

Figure 2. Mean values of respective r_1, r_2, Δ_m and Δ_f over 100 simulated data vs the number of slices, $3 \cdots 15$, in Model 2.

Table 1. Means and Standard Deviations of TPR and FPR, respectively, over 100 simulated data in Model 2.

	S-FT	S-SIR
TPR	0.8700(0.3380)	0.6300(0.4852)
FPR	0.0650(0.2418)	0.1207(0.3240)

four methods, indicating that FT is the best method for this model. We also conduct SVS for S-FT and S-SIR and report the respective TPR and FPR over 100 simulated data. The number of slices for S-SIR is 5 and the number of $\boldsymbol{\omega}$ for S-FT is 50. Table 1 shows that S-FT has larger TPR and smaller FPR compared to these of S-SIR, thus better results for S-FT than those of S-SIR.

Model 5.3: *This model is similar to Example* 4.1 *of* Bentler and Xie (2000). $Y = X_1 + 0.5\epsilon$, *with* $p = 4$, $d = 1$ *and* $B = e_1$. *Predictor vector* $\mathbf{X}_i$ *follows multivariate normal distribution with the mean* $(1, 2, 3, 4)$ *and equi-correlation matrix with a variance of 1 and a correlation of* 0.5, *and* $\epsilon \sim N(0, 1)$.

Table 2. Percentages of correctly detected dimensions in Model 3.

Sample Size	Weighted χ^2	Scale Stat.	Adj Stat.	SED
400	0.0000	1.0000	0.0000	0.9500
600	0.1800	1.0000	0.1700	1.0000
800	1.0000	1.0000	1.0000	1.0000

Table 3. TPR and FPR over 100 simulated data in Model 3.

	$n = 400$		$n = 600$		$n = 800$	
	S-FT	S-SIR	S-FT	S-SIR	S-FT	S-SIR
TPR	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
FPR	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

We check the three asymptotic dimension tests: the Weighted Chi-square Statistic, Scaled Statistic, and Adjusted Statistic, as well as the SED method (Zhu et al. 2010a) for this model. The size of $\boldsymbol{\omega}$ is 50, and we use three sample sizes of $n = 400,600,800$. The percentages of correctly detected dimensions among 100 simulated data are reported in Table 2, which shows that the Scaled Statistic performs better than Weighted Chi-square Statistic and Adjusted Statistic. This is consistent with Example 4.1 of Bentler and Xie (2000). The proportions of the correctly identified dimensions for the three asymptotic tests are 100% when the sample size is 800, resulting in more accurate estimates for larger sample sizes. Nevertheless, the Scaled test statistic is the best among all four tests. Moreover, we report TPR and FPR for S-FT and S-SIR, respectively, over 100 simulated data in Table 3, and the results are optimal.

Model 5.4: *This is Example 3 of* Zhu et al. (2010c). $Y_1 = 1 + \beta_1^{\mathrm{T}}\mathbf{X} + \sin(\beta_2^{\mathrm{T}}\mathbf{X}) + \epsilon_1$, $Y_2 = \beta_2^{\mathrm{T}}\mathbf{X}/(0.5 + (\beta_1^{\mathrm{T}}\mathbf{X} + 1)^2) + \epsilon_2$, $Y_3 = |\beta_1^{\mathrm{T}}\mathbf{X}|\epsilon_3$, $Y_4 = \epsilon_4$, $Y_5 = \epsilon_5$, *with* $p = 20$, $d = 2$, *and* $\beta_1 = e_1$ *and* $\beta_2 = e_2 + e_3$. *Predictor* $\mathbf{X}_i \sim N(0, I)$, $n = 2000$ *and* $\boldsymbol{\epsilon}_i = (\epsilon_1, \epsilon_2, \ldots, \epsilon_5)^{\mathrm{T}} \sim N_5(\mathbf{0}, \boldsymbol{\Sigma})$, *where* $\boldsymbol{\Sigma} = \begin{pmatrix} A & 0 \\ 0 & D \end{pmatrix}$, $A = \begin{pmatrix} 1 & -1/2 \\ -1/2 & 1/2 \end{pmatrix}$ *and* $D = \mathrm{diag}(1/2, 1/3, 1/4)$.

This is a multivariate model. Figure 3 plots mean values of respective r_1, r_2, Δ_m and Δ_f over 100 simulated data vs sizes of $\boldsymbol{\omega}$: $\{10, 20, \ldots, 100\}$. All four criteria show that FT estimates tend to improve as the size of $\boldsymbol{\omega}$ increases and then become stable. On the other hand, Zhu et al. (2010c) has demonstrated the advantage of FT for multivariate regression over other methods: the projective resampling method (Li et al. 2008), the K-means inverse regression (Setodji and Cook 2004), alternative SIR (Li et al. 2005), nearest neighbour inverse regression (Hsing 1999) and moment approach (Yin and Bura 2006). We omit the related comparisons here.

The left panel of Figure 4 shows the Weighted Chi-square Statistic, Scaled Statistic, Adjusted Statistic and SED test. Compared with the Weighted Chi-square Statistic and Adjusted Statistic, the Scaled Statistic is better. (Actually, we also use sample sizes $n = 1000$, but not reported here. The Scaled Statistic still performs well, which indicates the Scaled Statistic converges more quickly.) If the size of $\boldsymbol{\omega}$ is large enough, the performance of the Scaled Statistic becomes stable, and the proportion of correct decisions gets closer to 1, which agrees with the estimation accuracy. The Scaled Statistic is better than SED when the size of $\boldsymbol{\omega}$ is large enough. However, SED is not stable. When the size of $\boldsymbol{\omega}$ is large enough

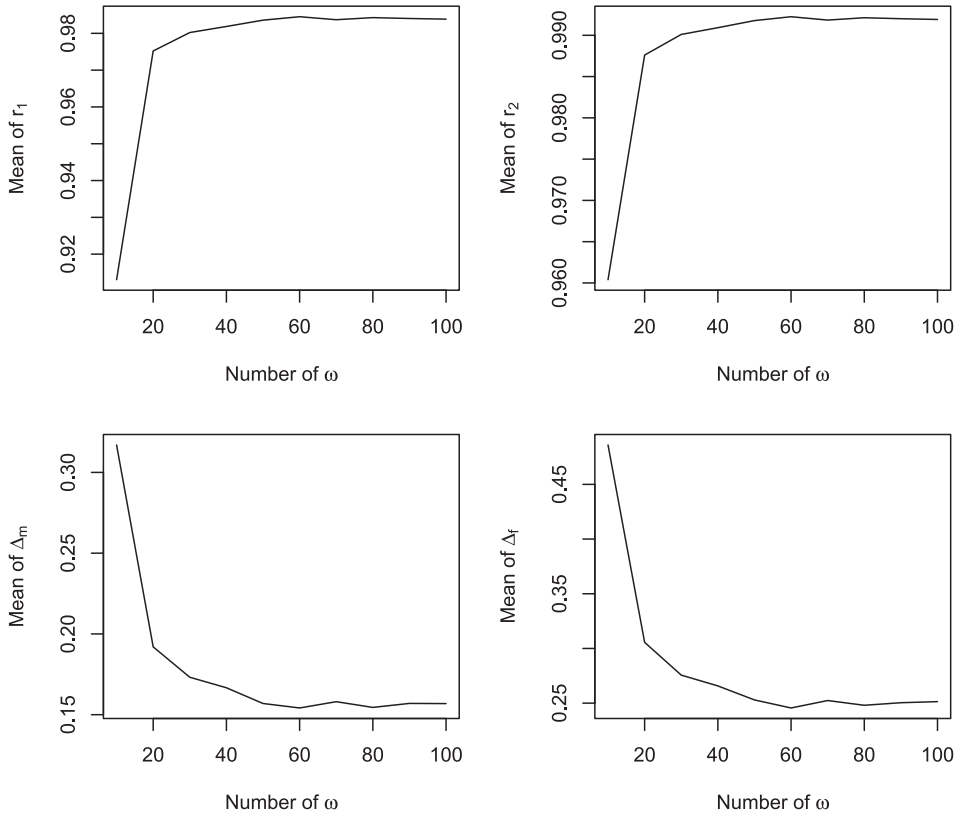

Figure 3. Mean values of respective r_1, r_2, Δ_m and Δ_f over 100 simulated data vs sizes of $\boldsymbol{\omega}$: $\{10, 20, \ldots, 100\}$ in Model 4.

(over 60 as in Figure 4), its result is worse, contradicting the accuracy of the estimate. The middle and right panels of Figure 4 show TPR and FPR, respectively, for S-FT and S-SIR. TPR values are similar for the two methods with smaller FPR for S-FT when the size of $\boldsymbol{\omega}$ is between 20 to 80 compared to S-SIR. Regardless, FPRs are all relatively small using either S-SIR or S-FT.

Additionally, we change the number of predictors to be $p = \{10, 20\}$ and use sample sizes $n = \{1000, 2000\}$, with the number of response variables to be $q = \{5, 10, 15\}$ (not reported here). The number of predictors and sample size affect the asymptotic results in testing the dimension. As sample size increases, the performance of the Weighted Chi-square Statistic, Scaled Statistic, and Adjusted Statistic improve, especially for the Scaled Statistic. If the number of predictors increases, a larger sample size is needed for asymptotic results to converge. While adding some noise response variables and changing the number of response variables does not significantly affect the results.

Model 5.5: *For* $W = 0$, *let* $Y = X_1 + 0.1\epsilon$, *with* $B_1 = e_1$. *Predictors* $\mathbf{X}_{i1} \sim \frac{1}{4}N_p(\boldsymbol{\mu}_1, \Sigma_1) + \frac{1}{2}N_p(\boldsymbol{\mu}_2, \Sigma_2) + \frac{1}{4}N_p(\boldsymbol{\mu}_3, \Sigma_3)$, *where* $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_3 = (1, 0, \ldots, 0)^{\mathrm{T}}$, $\boldsymbol{\mu}_2 = (2, 0, \ldots, 0)^{\mathrm{T}}$, $\Sigma_1 = \Sigma_2 = \sqrt{0.1}I_p$ *and* $\Sigma_3 = \sqrt{10}I_p$. *Let* $p = 10$, *and* ϵ *is a uniform* (0,1), *with 1000 observations. For* $W = 1$, *let* Y *be the* Y_2 *in the model 4 with* $B_2 = e_2 + e_3$ *and 1000 observations.*

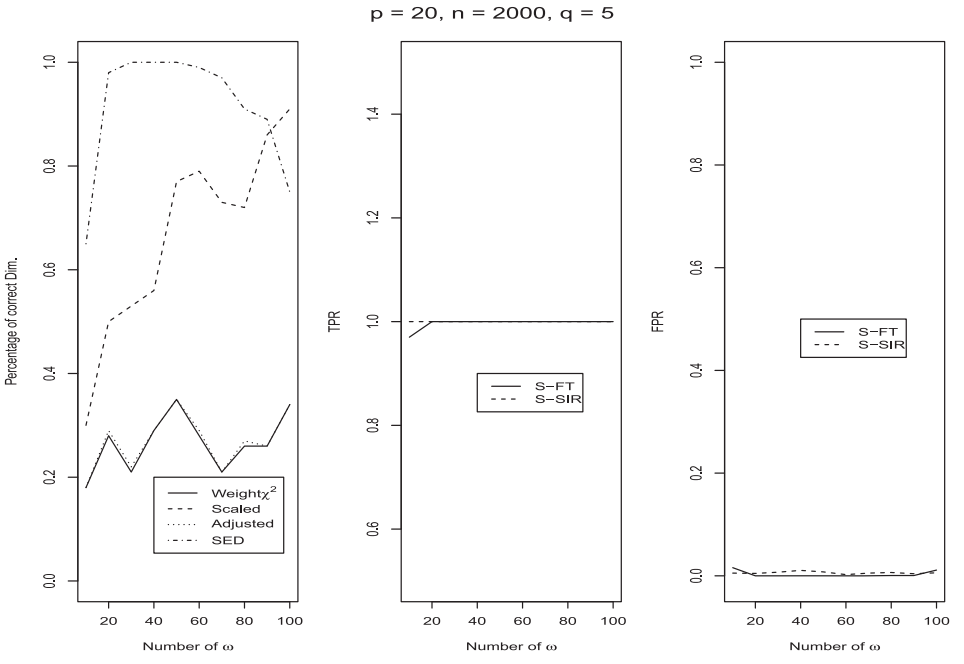

Figure 4. Left panel: percentages of correctly detected dimension over 100 simulated data vs sizes of $\boldsymbol{\omega}$: {10, 20, . . . , 100} in Model 4; Middle and Right panel: TPR and FPR over 100 simulated data vs sizes of $\boldsymbol{\omega}$.

Table 4. Mean values of respective r_1, r_2, Δ_m and Δ_f over 100 simulated data for PSIR and PFT and proportion of correctly detected dimension in Model 5.

	r_1	r_2	Δ_m	Δ_f	$Proportion_c$
PSIR	0.9889	0.9944	0.1249	0.2081	0.9700
PFT	0.9930	0.9965	0.1066	0.1640	1.0000

This example compares PSIR and PFT. The number of slices for PSIR is 10, and the size of $\boldsymbol{\omega}$ for PFT is 50. We replicate 100 times for the model and then calculate the averages of respective r_1, r_2, Δ_m and Δ_f and the proportion of correctly detected dimensions using the Scaled Statistics, say, $Proportion_c$. Table 4 shows that PFT performs consistently better than PSIR does in every criterion.

Model 5.6: *This is Model 4, except:* $\beta_1 = e_1 + e_2 + e_3 + e_4, \beta_2 = e_5 + e_6 + \cdots + e_{12}, p = 1000$, *and* $n = 400$. *This is a large p, small n problem.*

We use path I algorithm in Section 4. We use 10 slices for SIR, 50 sizes of $\boldsymbol{\omega}$ for FT and $p_1 = 15$ as the number of predictors in each step for both SIR and FT. The asymptotic Scaled Statistic test is used in each step for estimating the dimensionality. Table 5 reports the average values for each criterion over 100 simulated data. SSDR-FT performs consistently better than that of SSDR-SIR, except that both TPR and FPR are the same for the two methods.

Table 5. Accuracy for large p, small n data in Model 6.

	Corr1	Corr2	Δ_f	Δ_m	TPR	FPR
SSDR-SIR	0.8764	0.8024	0.7264	0.4567	0.9783	0.0134
	(0.2034)	(0.1852)	(0.1058)	(0.0788)	(0.1062)	(0.0646)
SSDR-FT	0.9565	0.9106	0.6335	0.3912	0.9783	0.0134
	(0.0218)	(0.0350)	(0.1018)	(0.0705)	(0.1062)	(0.0646)

5.2. Data analysis

The data set is the '2015 Planning Database' (PDB) with 2010 Census and 2009–2013 American Community Survey data, which is publicly available (http://goo.gl/LlcwY7). PDB assembles information from housing, demographic, socioeconomic, and Census operational data, and accumulates at the block-group level. A census block is the smallest geographic unit used by the Census Bureau, and a block-group comprises multiple blocks, usually containing between 600 and 3,000 people. The PDB comprises approximately 220,000 block groups.

The response variable is the number of people with one type of health insurance coverage (Y). A total of 15 variables are identified as relevant candidate predictor variables. See Table 6. Because most of the variables are count numbers with a large range of values, we treat all of them as continuous variables.

Table 6. The first two columns are the variable numbers and definitions as given in the 2015 Planning Database documentation. The last column is the variable notation used in our manuscript.

No.	Definitions	Notation (Box-Cox)
73	Number of people ages 25 years and over at the time of interview with a college degree or higher in the ACS population	$X_1((X_1+0.5)^{0.33})$
77	Number of people classified as below the poverty level given their total family income within the last year, family size, and family composition in the ACS population	$X_2((X_2+0.5)^{0.33})$
103	Number of ACS households in which the householder and his or her spouse are listed as members of the same household; does not include same-sex married couples	$X_3((X_3+0.5)^{0.57})$
112	Number of ACS households where a householder lives alone or with nonrelatives only; includes same-sex couples where no relatives of the householder are present	$X_4((X_4+0.5)^{0.33})$
115	Number of ACS households where a householder lives alone	$X_5((X_5+0.5)^{0.4})$
124	Number of ACS families with related children under 6 years old	$X_6((X_6+0.5)^{0.5})$
130	Median ACS household income for the block group	$X_7((X_7+0.5)^{0.33})$
132	Median ACS household income for the tract	$X_8((X_8+0.5)^{0.16})$
145	Number of 2010 Census occupied housing units that are not owner occupied, whether they are rented or occupied without payment of rent	$X_9((X_9+0.5)^{0.27})$
149	Number of ACS housing units where owner or co-owner lives in it	$X_{10}((X_{10}+0.5)^{0.5})$
151	Number of ACS housing units in which the structure contains only that single unit	$X_{11}((X_{11}+0.5)^{0.5})$
153	Number of ACS housing units in which the structure contains 2 or more housing units	$X_{12}((X_{12}+0.5)^{0.33})$
155	Number of ACS housing units in which the structure contains 10 or more housing units	$X_{13}((X_{13}+0.5)^{-0.06})$
167	Median of ACS respondents' house value estimates for the block group	$X_{14}((X_{14}+0.5)^{0.15})$
169	Median of ACS respondents' house value estimates for the tract	$X_{15}((X_{15}+0.5)^{0.1})$
3	Name of State or statistically equivalent territory; island territories are excluded from this analysis; these values are converted to a categorical variable based on 9 Census-designated geographical regions; only used for partial SIR analysis	W
79	Number of people with one type of health insurance coverage in the ACS	Y

We focus on the block groups in Rhode Island, which have 4270 blocks. We first excluded any observations where the variables had missing values. There were 4098 blocks left for Rhode Island. We then used Box-Cox transformation for the predictors to ensure that the linearity condition was approximately satisfied. Transformation for each variable is in last column of Table 6 (inside the parenthesis).

Using the Scaled Statistic for all the blocks in Rhode Island, the estimated dimension (using 50 as the size of ω) is one. In addition, if we plot the scatter plot (Figure 5) of response variable versus the first reduced variable, we can see the strong association. The second reduced variable does not contribute much. Hence, one dimension is sufficient to capture the CS. Thus, we used one dimension for the following analysis.

To illustrate the advantages of FT, we used five datasets: the first 100 blocks, the first 200 blocks, the first 400 blocks, the first 800 blocks, and all blocks of Rhode Island. For each data, we estimated the vector $\hat{\beta}$ (of the CS). We then bootstrapped 100 samples from that data and obtained the bootstrap estimate $\hat{\beta}^b$ for each bootstrap sample. Then we compare means of r_2 between the bootstrap estimate $\hat{\beta}^b$ and $\hat{\beta}$ using the following methods: FT, SIR, SAVE, PHD, FIRE, and DIRE. For SIR and SAVE, we fix the number of slices to be 5, which is typically what researchers suggested. For FIRE and DIRE, the sequence of slices is $\{3, 4, 5, \ldots, 15\}$, which is what Cook and Zhang (2014) suggested. Table 7 shows the results. It indicates that when sample size increases, every method performs better, which is expected. However, none of them is comparable with FT, until sample size reaches to 4098. On the other hand, FT approach provides the most accurate and stable estimates among all these methods and across all sample sizes. Even in the small sample size of 100, FT still provides an accurate estimate with $r_2 = 0.9840$. It indicates that its estimates converge much faster than all other methods.

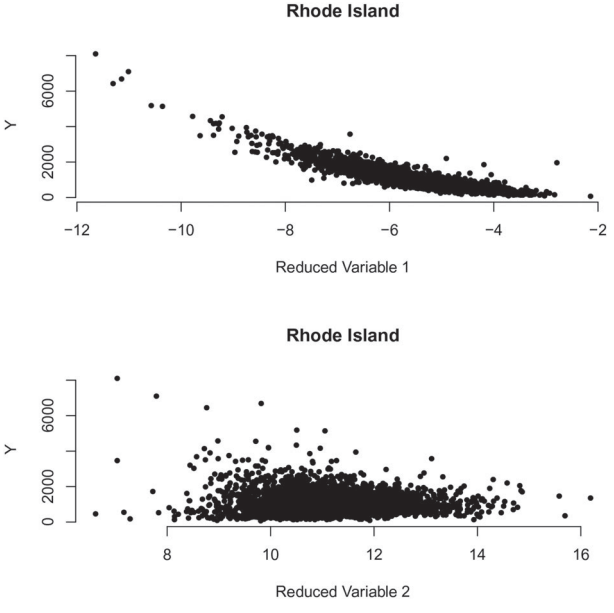

Figure 5. Scatter Plots: response variable versus the first and the second reduced variable.

Table 7. Means and standard deviations of r_2 for each method: FT, SIR, SAVE, PHD, FIRE and DIRE over sample sizes: {100, 200, 400, 800, 4098}.

	Rhode	Rhode	Rhode	Rhode	Rhode
r_2	($n = 100$)	($n = 200$)	($n = 400$)	($n = 800$)	($n = 4098$)
FT	0.9840	0.9879	0.9926	0.9956	0.9991
	(0.0079)	(0.0058)	(0.0034)	(0.002)	(4e-04)
SIR	0.5543	0.7072	0.8591	0.892	0.9754
	(0.2334)	(0.2209)	(0.1233)	(0.059)	(0.0147)
SAVE	0.4136	0.6017	0.7417	0.7319	0.9629
	(0.2676)	(0.2834)	(0.2612)	(0.2984)	(0.0281)
PHD	0.7665	0.6128	0.7787	0.7944	0.8597
	(0.2437)	(0.3054)	(0.1926)	(0.2355)	(0.1156)
FIRE	0.4857	0.5056	0.8296	0.9133	0.9869
	(0.2424)	(0.2954)	(0.1392)	(0.0500)	(0.0082)
DIRE	0.3816	0.3669	0.6911	0.9002	0.9882
	(0.2096)	(0.2157)	(0.1600)	(0.0561)	(0.0066)

6. Discussion

Using FT, we develop a complete package for estimating CS. We provide its estimator, algorithm and asymptotic properties. It is important to note that FT approach avoids the trouble of selecting the number of slices in inverse regression and provides a natural solution for multivariate response. We further extended this approach to partial SDR, SVS, and large p, small n data. Given the current FT approach, a general discussion about inverse regression family may be developed. Such an investigation is our on-going project.

Acknowledgments

The authors would like to thank the Editor, an Associate Editor and a referee for their valuable comments and suggestions, which lead to a greatly improved paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by an NSF [grant CIF-1813330].

References

Aragon, Y. (1997), 'A Gauss Implementation of Multivariate Sliced Inverse Regression', *Computational Statistics*, 12, 355–372.

Bentler, P.M., and Xie, J. (2000), 'Corrections to Test Statistics in Principal Hessian Directions', *Statistics and Probability Letters*, 47, 381–389.

Chiaromonte, F., Cook, R.D., and Li, B. (2002), 'Sufficient Dimension Reduction in Regressions with Categorical Predictors', *The Annals of Statistics*, 30, 475–497.

Cook, R.D. (1996), 'Graphics for Regressions with a Binary Response', *Journal of the American Statistical Association*, 91, 983–992.

Cook, R.D. (1998), *Regression Graphics: Ideas for Studying Regressions through Graphics*, New York, NY: John Wiley & Sons Inc.

Cook, R.D., and Li, B. (2002), 'Dimension Reduction for Conditional Mean in Regression', *The Annals of Statistics*, 30, 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 Setodji, C.M. (2003), 'A Model-free Test for Reduced Rank in Multivariate Regression', *Journal of the American Statistical Association*, 98, 340–351.
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.
Cook, R.D., and Zhang, X. (2014), 'Fused Estimators of the Central Subspace in Sufficient Dimension Reduction', *Journal of the American Statistical Association*, 109, 815–827.
Hsing, T. (1999), 'Nearest Neighbor Inverse Regression', *The Annals of Statistics*, 27, 697–731.
Hsing, T., and Carroll, R.J. (1992), 'An Asymptotic Theory for Sliced Inverse Regression', *The Annals of Statistics*, 20, 1040–1061.
Li, K.C. (1991), 'Sliced Inverse Regression for Dimension Reduction (with discussion)', *Journal of the American Statistical Association*, 86, 316–327.
Li, K.C. (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, L. (2007), 'Sparse Sufficient Dimension Reduction', *Biometrika*, 94, 603–613.
Li, B., Cook, R.D., and Chiaromonte, F. (2003), 'Dimension Reduction for Conditional Mean in Regressions with Categorical Predictors', *The Annals of Statistics*, 31, 1636–1668.
Li, B., Zha, H., and Chiaromonte, F. (2005), 'Contour Regression: A General Approach to Dimension Reduction', *The Annals of Statistics*, 33, 1580–1616.
Li, B., Wen, S., and Zhu, L.X. (2008), 'On a Projective Resampling Method for Dimension Reduction with Multivariate Responses', *Journal of the American Statistical Association*, 103, 1177–1186.
Li, R., Zhong, W., and Zhu, L.P. (2012), 'Feature Screening via Distance Correlation Learning', *Journal of the American Statistical Association*, 107, 1129–1139.
Luo, W., Li, B., and Yin, X. (2014), 'On Efficient Dimension Reduction with Respect to a Statistical Functional of Interest', *The Annals of Statistics*, 42, 382–412.
Saracco, J. (2005), 'Asymptotic for Pooled Marginal Slicing Estimator Based on SIRα Approach', *Journal of Multivariate Analysis*, 96, 117–135.
Setodji, C.M., and Cook, R.D. (2004), 'K-means Inverse Regression', *Technometrics*, 46, 421–429.
Ye, Z., and Weiss, R.E. (2003), 'Using the Bootstrap to Select One of a New Class of Dimension Reduction Methods', *Journal of the American Statistical Association*, 98, 968–979.
Yin, X., and Bura, E. (2006), 'Moment-based Dimension Reduction for Multivariate Response Regression', *Journal of Statistical Planning and Inference*, 136, 3675–3688.
Yin, X., and Cook, R.D. (2002), 'Dimension Reduction for the Conditional *k*th Moment in Regression', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 159–175.
Yin, X., and Hilafu, H. (2015), 'Sequential Sufficient Dimension Reduction for Large p, Small n Problems', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77, 879–892.
Yin, X., and Li, B. (2011), 'Sufficient Dimension Reduction Based on an Ensemble of Minimum Average Variance Estimators', *The Annals of Statistics*, 39, 3392–3416.
Yin, X., Li, B., and Cook, R.D. (2008), 'Successive Direction Extraction for Estimating the Central Subspace in a Multiple-index Regression', *Journal of Multivariate Analysis*, 99, 1733–1757.
Zhu, L.X., and Ng, K.W. (1995), 'Asymptotic of Sliced Inverse Regression', *Statistica Sinica*, 5, 727–736.
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.
Zhu, L.P., and Zhu, L.X. (2009), 'Dimension Reduction for Conditional Variance in Regressions', *Statistica Sinica*, 19, 869–883.
Zhu, L.P., Yu, Z., and Zhu, L.X. (2010a), 'A Sparse Eigen-decomposition Estimation in Semiparametric Regression', *Computational Statistics & Data Analysis*, 54, 976–986.
Zhu, L.P., Zhu, L.X., and Feng, Z.H. (2010b), 'Dimension Reduction in Regressions through cumulative Slicing Estimation', *Journal of the American Statistical Association*, 105, 1455–1466.
Zhu, L.P., Zhu, L.X., and Wen, S.Q. (2010c), 'On Dimension Reduction in Regressions with Multivariate Responses', *Statistica Sinaca*, 20, 1291–1307.

Appendix

Proof of Equivalent of FT and SIR when response variable is categorical: Assume Y is univariate, and Y has K levels $\{0, 1, \ldots, K-1\}$ with probability $P_y = P(Y = y) > 0, y \in \{0, 1, \ldots, K-1\}$. Let $\mathcal{S}_{ft} = \text{Span}\{\psi(\omega), \omega \in \mathbb{R}\}$ and $\mathcal{S}_{sir} = \text{Span}\{\text{E}(\mathbf{Z} \mid Y = y), y \in \{0, 1, \ldots, K-1\}\}$.

$$\begin{aligned}\psi(\omega) &= \text{E}[\text{E}(\text{e}^{\text{i}\omega y}\mathbf{Z} \mid Y = y)] \\ &= \text{E}(\mathbf{Z} \mid Y = 0)P(Y = 0) + \text{E}(\text{e}^{\text{i}\omega}\mathbf{Z} \mid Y = 1)P(Y = 1) + \cdots \\ &\quad + \text{E}(\text{e}^{\text{i}\omega(K-1)}\mathbf{Z} \mid Y = K-1)P(Y = K-1) \\ &= P_0\text{E}(\mathbf{Z} \mid Y = 0) + P_1\,\text{e}^{\text{i}\omega}\text{E}(\mathbf{Z} \mid Y = 1) + \cdots + P_{K-1}\,\text{e}^{\text{i}\omega(K-1)}\text{E}(\mathbf{Z} \mid Y = K-1).\end{aligned}$$

Because $\text{E}(\mathbf{Z} \mid Y) \in \mathcal{S}_{sir}$, then $\mathcal{S}_{ft} \subseteq \mathcal{S}_{sir}$.

Now, choose $\omega_1, \ldots, \omega_{K-1}$ such that they are all different numbers.

$$\begin{aligned}\psi(0) &= P_0\text{E}(\mathbf{Z} \mid Y = 0) + P_1\text{E}(\mathbf{Z} \mid Y = 1) + \cdots + P_{K-1}\text{E}(\mathbf{Z} \mid Y = K-1), \\ \psi(\omega_1) &= P_0\text{E}(\mathbf{Z} \mid Y = 0) + P_1\,\text{e}^{\text{i}\omega_1}\text{E}(\mathbf{Z} \mid Y = 1) + \cdots + P_{K-1}\,\text{e}^{\text{i}\omega_1(K-1)}\text{E}(\mathbf{Z} \mid Y = K-1), \\ &\vdots \\ \psi(\omega_{K-1}) &= P_0\text{E}(\mathbf{Z} \mid Y = 0) + P_1\,\text{e}^{\text{i}\omega_{K-1}}\text{E}(\mathbf{Z} \mid Y = 1) + \cdots + P_{K-1}\,\text{e}^{\text{i}\omega_{K-1}(K-1)}\text{E}(\mathbf{Z} \mid Y = K-1).\end{aligned}$$

And the following matrix is nonsingular:

$$A = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \text{e}^{\text{i}\omega_1} & \text{e}^{\text{i}2\omega_1} & \cdots & \text{e}^{\text{i}(K-1)\omega_1} \\ 1 & \text{e}^{\text{i}\omega_2} & \text{e}^{\text{i}2\omega_2} & \cdots & \text{e}^{\text{i}(K-1)\omega_2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & \text{e}^{\text{i}\omega_{K-1}} & \text{e}^{\text{i}2\omega_{K-1}} & \cdots & \text{e}^{\text{i}(K-1)\omega_{K-1}} \end{pmatrix} \text{diag}(P_y).$$

Because $|A| = \prod_{y=1}^{K-1} P_y \prod_{y=2}^{K-1}(\text{e}^{\text{i}\omega_y} - \text{e}^{\text{i}\omega_1}) \prod_{y=3}^{K-1}(\text{e}^{\text{i}\omega_y} - \text{e}^{\text{i}\omega_2}) \cdots (\text{e}^{\text{i}\omega_{K-1}} - \text{e}^{\text{i}\omega_{K-2}}) \neq 0$, then we have $(\text{E}(\mathbf{Z} \mid Y = 0), \ldots, \text{E}(\mathbf{Z} \mid Y = K-1))^{\text{T}} = A^{-1}(\psi(0), \psi(\omega_1), \ldots \psi(\omega_{K-1}))^{\text{T}}$. Because $\psi(0)$, $\psi(\omega_1), \ldots, \psi(\omega_{K-1}) \in \mathcal{S}_{ft}$, then $\text{E}(\mathbf{Z} \mid Y = y) \in \mathcal{S}_{ft}$ for $y \in \{0, 1, \ldots, K-1\}$. That is, $\mathcal{S}_{sir} \subseteq \mathcal{S}_{ft}$. Hence, $\mathcal{S}_{ft} = \mathcal{S}_{sir}$. ■

Proof of Proposition 2.2***:*** To obtain the asymptotic distribution of $\hat{\Lambda}_d$, fix t and choose $\{\boldsymbol{\omega}_j\}_{j=1}^t$. For $j = 1, \ldots, t$, define:

$$\hat{\boldsymbol{\psi}}_{j1} = \frac{1}{n}\sum_{k=1}^{n} \hat{\mathbf{z}}_k \cos(\boldsymbol{\omega}_j^{\text{T}}\mathbf{y}_k), \quad \boldsymbol{\psi}_{j1} = \text{E}[\mathbf{Z}\cos(\boldsymbol{\omega}_j^{\text{T}}\mathbf{Y})],$$

$$\hat{\boldsymbol{\psi}}_{j2} = \frac{1}{n}\sum_{k=1}^{n} \hat{\mathbf{z}}_k \sin(\boldsymbol{\omega}_j^{\text{T}}\mathbf{y}_k), \quad \boldsymbol{\psi}_{j2} = \text{E}[\mathbf{Z}\sin(\boldsymbol{\omega}_j^{\text{T}}\mathbf{Y})].$$

Let $\hat{\Psi} = (\hat{\boldsymbol{\psi}}_{11}, \hat{\boldsymbol{\psi}}_{12}, \ldots, \hat{\boldsymbol{\psi}}_{t1}, \hat{\boldsymbol{\psi}}_{t2})$ and $\Psi = (\boldsymbol{\psi}_{11}, \boldsymbol{\psi}_{12}, \ldots, \boldsymbol{\psi}_{t1}, \boldsymbol{\psi}_{t2})$. Following Cook (1998, p. 207), by Singular-Value Decomposition, $\Psi = \Gamma^{\text{T}}\begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}\Phi$, where Γ and Φ are respective $p \times p$ and $2t \times 2t$ orthogonal matrices, and D is a $d \times d$ diagonal matrix of positive values. Let $\Gamma^{\text{T}} = (\Gamma_1, \Gamma_0)$ and $\Phi^{\text{T}} = (\Phi_1, \Phi_0)$, where Γ_0 is $p \times (p-d)$ and Φ_0 is $2t \times (2t-d)$. In X-scale and $j = 1, \ldots, t$,

define:

$$\hat{\boldsymbol{\theta}}_{j1} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k \cos(\boldsymbol{\omega}_j^{\mathrm{T}}\mathbf{y}_k), \quad \boldsymbol{\theta}_{j1} = \mathrm{E}[\mathbf{X}\cos(\boldsymbol{\omega}_j^{\mathrm{T}}\mathbf{Y})],$$

$$\hat{\boldsymbol{\theta}}_{j2} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k \sin(\boldsymbol{\omega}_j^{\mathrm{T}}\mathbf{y}_k), \quad \boldsymbol{\theta}_{j2} = \mathrm{E}[\mathbf{X}\sin(\boldsymbol{\omega}_j^{\mathrm{T}}Y)].$$

$$\hat{\Theta} = (\hat{\boldsymbol{\theta}}_{11}, \hat{\boldsymbol{\theta}}_{12}, \ldots, \hat{\boldsymbol{\theta}}_{t1}, \hat{\boldsymbol{\theta}}_{t2}), \quad \Theta = (\boldsymbol{\theta}_{11}, \boldsymbol{\theta}_{12}, \ldots, \boldsymbol{\theta}_{t1}, \boldsymbol{\theta}_{t2}).$$

$$V = \Psi\Psi^{\mathrm{T}}, \quad \hat{V} = \hat{\Psi}\hat{\Psi}^{\mathrm{T}}.$$

Set $\hat{Q} = ((1/n)\sum\cos(\boldsymbol{\omega}_1^{\mathrm{T}}\mathbf{y}_k), (1/n)\sum\sin(\boldsymbol{\omega}_1^{\mathrm{T}}\mathbf{y}_k), \ldots, (1/n)\sum\cos(\boldsymbol{\omega}_t^{\mathrm{T}}\mathbf{y}_k), (1/n)\sum\sin(\boldsymbol{\omega}_t^{\mathrm{T}}\mathbf{y}_k))^{\mathrm{T}}$, and $Q = (\mathrm{E}(\cos\boldsymbol{\omega}_1^{\mathrm{T}}\mathbf{Y}), \mathrm{E}(\sin\boldsymbol{\omega}_1^{\mathrm{T}}\mathbf{Y}), \ldots, \mathrm{E}(\cos\boldsymbol{\omega}_t^{\mathrm{T}}\mathbf{Y}), \mathrm{E}(\sin\boldsymbol{\omega}_t^{\mathrm{T}}\mathbf{Y}))^{\mathrm{T}}$.

Look at the one column of Ψ as an example, say, $\mathrm{E}[\mathbf{Z}\cos(\boldsymbol{\omega}^{\mathrm{T}}\mathbf{Y})]$. Then

$$\begin{aligned}\mathrm{E}[\mathbf{Z}\cos(\boldsymbol{\omega}^{\mathrm{T}}\mathbf{Y})] &= \mathrm{E}[\Sigma^{-1/2}(\mathbf{X}-\boldsymbol{\mu})\cos(\boldsymbol{\omega}^{\mathrm{T}}\mathbf{Y})] \\ &= \Sigma^{-1/2}\mathrm{E}[(\mathbf{X}-\boldsymbol{\mu})\cos(\boldsymbol{\omega}^{\mathrm{T}}\mathbf{Y})] \\ &= \Sigma^{-1/2}\{\mathrm{E}[\mathbf{X}\cos(\boldsymbol{\omega}^{\mathrm{T}}\mathbf{Y})] - \boldsymbol{\mu}\mathrm{E}[\cos(\boldsymbol{\omega}^{\mathrm{T}}\mathbf{Y})]\}.\end{aligned}$$

So $\Psi = \Sigma^{-1/2}\Theta - \Sigma^{-1/2}\boldsymbol{\mu}Q^{\mathrm{T}}$. And $\Gamma_0^{\mathrm{T}}\Psi\Phi_0 = 0$, that is, $\Gamma_0^{\mathrm{T}}\Sigma^{-1/2}(\Theta - \boldsymbol{\mu}Q^{\mathrm{T}})\Phi_0 = 0$. Define $\hat{A} = \hat{\Sigma}^{-1/2}\Sigma^{1/2}$, then

$$\begin{aligned}\sqrt{n}\Gamma_0^{\mathrm{T}}\hat{\Psi}\Phi_0 &= \sqrt{n}\Gamma_0^{\mathrm{T}}\hat{\Sigma}^{-1/2}(\hat{\Theta} - \bar{\boldsymbol{x}}\hat{Q}^{\mathrm{T}})\Phi_0 = \sqrt{n}\Gamma_0^{\mathrm{T}}\hat{A}\Sigma^{-1/2}(\hat{\Theta} - \bar{\boldsymbol{x}}\hat{Q}^{\mathrm{T}})\Phi_0 \\ &= \sqrt{n}\Gamma_0^{\mathrm{T}}(\hat{A} - I + I)\Sigma^{-1/2}[\hat{\Theta} - \Theta + \Theta - \boldsymbol{\mu}Q^{\mathrm{T}} + \boldsymbol{\mu}(Q^{\mathrm{T}} - \hat{Q}^{\mathrm{T}}) + (\boldsymbol{\mu} - \bar{\boldsymbol{x}})\hat{Q}^{\mathrm{T}}]\Phi_0.\end{aligned}$$

Here $(\hat{A} - I)\Sigma^{-1/2}(\hat{\Theta} - \Theta) = O_p(1/n)$, $(\hat{A} - I)\Sigma^{-1/2}\boldsymbol{\mu}(Q^{\mathrm{T}} - \hat{Q}^{\mathrm{T}}) = O_p(1/n)$, $(\hat{A} - I)\Sigma^{-1/2}(\boldsymbol{\mu} - \bar{\boldsymbol{x}})\hat{Q}^{\mathrm{T}} = O_p(1/n)$ and $\Gamma_0^{\mathrm{T}}(\hat{A} - I)\Sigma^{-1/2}(\Theta - \boldsymbol{\mu}Q^{\mathrm{T}})\Phi_0 = 0$. Hence,

$$\begin{aligned}\sqrt{n}\Gamma_0^{\mathrm{T}}\hat{\Psi}\Phi_0 &= \sqrt{n}\Gamma_0^{\mathrm{T}}\Sigma^{-1/2}[\hat{\Theta} - \Theta + \boldsymbol{\mu}(Q^{\mathrm{T}} - \hat{Q}^{\mathrm{T}}) + (\boldsymbol{\mu} - \bar{\boldsymbol{x}})Q^{\mathrm{T}}]\Phi_0 + O_p\left(\frac{1}{n}\right) \\ &= \sqrt{n}\Gamma_0^{\mathrm{T}}\Sigma^{-1/2}[\hat{\Theta} - \Theta + \boldsymbol{\mu}(Q - \hat{Q})^{\mathrm{T}} + (\boldsymbol{\mu} - \bar{\boldsymbol{x}})Q^{\mathrm{T}}]\Phi_0 + O_p\left(\frac{1}{n}\right).\end{aligned}$$

By central limit theorem, we have

$$\sqrt{n}((\mathrm{vec}(\hat{\Theta} - \Theta))^{\mathrm{T}}, (\hat{Q} - Q)^{\mathrm{T}}, (\bar{\boldsymbol{x}} - \boldsymbol{\mu})^{\mathrm{T}})^{\mathrm{T}} \to N_{2pt+2t+p}\left(\mathbf{0}, \tau = \begin{pmatrix} \Delta_{xy} & \Delta_{xy,y} & \Delta_{xy,x} \\ \Delta_{xy,y}^{\mathrm{T}} & \Delta_y & \Delta_{y,x} \\ \Delta_{xy,x}^{\mathrm{T}} & \Delta_{y,x}^{\mathrm{T}} & \Sigma \end{pmatrix}\right),$$

where the τ will be defined as follows: $\mathrm{Cov}(\mathbf{X}\cos(\boldsymbol{\omega}_j^{\mathrm{T}}\mathbf{Y}), \mathbf{X}\cos(\boldsymbol{\omega}_k^{\mathrm{T}}\mathbf{Y})) = \Delta_{xy}^{j1,k1}$,

$$\mathrm{Cov}(\mathbf{X}\cos(\boldsymbol{\omega}_j^{\mathrm{T}}\mathbf{Y}), \mathbf{X}\sin(\boldsymbol{\omega}_k^{\mathrm{T}}\mathbf{Y})) = \Delta_{xy}^{j1,k2}, \quad \mathrm{Cov}(\mathbf{X}\sin(\boldsymbol{\omega}_j^{\mathrm{T}}\mathbf{Y}), \mathbf{X}\sin(\boldsymbol{\omega}_k^{\mathrm{T}}\mathbf{Y})) = \Delta_{xy}^{j2,k2},$$

$$\mathrm{Cov}(\mathbf{X}\cos(\boldsymbol{\omega}_j^{\mathrm{T}}\mathbf{Y}), \cos(\boldsymbol{\omega}_k^{\mathrm{T}}\mathbf{Y})) = \Delta_{xy,y}^{j1,k1}, \quad \mathrm{Cov}(\mathbf{X}\cos(\boldsymbol{\omega}_j^{\mathrm{T}}\mathbf{Y}), \sin(\boldsymbol{\omega}_k^{\mathrm{T}}\mathbf{Y})) = \Delta_{xy,y}^{j1,k2},$$

$$\mathrm{Cov}(\mathbf{X}\sin(\boldsymbol{\omega}_j^{\mathrm{T}}\mathbf{Y}), \sin(\boldsymbol{\omega}_k^{\mathrm{T}}\mathbf{Y})) = \Delta_{xy,y}^{j2,k2}, \quad \mathrm{Cov}(\mathbf{X}\cos(\boldsymbol{\omega}_j^{\mathrm{T}}\mathbf{Y}), \mathbf{X}) = \Delta_{xy,x}^{j1},$$

$$\mathrm{Cov}(\mathbf{X}\sin(\boldsymbol{\omega}_j^{\mathrm{T}}\mathbf{Y}), \mathbf{X}) = \Delta_{xy,x}^{j2}, \quad \mathrm{Cov}(\cos(\boldsymbol{\omega}_j^{\mathrm{T}}\mathbf{Y}), \cos(\boldsymbol{\omega}_k^{\mathrm{T}}\mathbf{Y})) = \Delta_y^{j1,k1},$$

$$\mathrm{Cov}(\cos(\boldsymbol{\omega}_j^{\mathrm{T}}\mathbf{Y}), \sin(\boldsymbol{\omega}_k^{\mathrm{T}}\mathbf{Y})) = \Delta_y^{j1,k2}, \quad \mathrm{Cov}(\sin(\boldsymbol{\omega}_j^{\mathrm{T}}\mathbf{Y}), \sin(\boldsymbol{\omega}_k^{\mathrm{T}}\mathbf{Y})) = \Delta_y^{j2,k2},$$

$$\mathrm{Cov}(\cos(\boldsymbol{\omega}_j^{\mathrm{T}}\mathbf{Y}), \mathbf{X}) = \Delta_{y,x}^{j1}, \quad \mathrm{Cov}(\sin(\boldsymbol{\omega}_j^{\mathrm{T}}\mathbf{Y}), \mathbf{X}) = \Delta_{y,x}^{k2},$$

$$\Delta_{xy} = \begin{array}{c} \\ p \\ p \\ \vdots \\ p \\ p \end{array} \begin{array}{c} \begin{array}{ccccc} p & p & \dots & p & p \end{array} \\ \begin{pmatrix} \Delta_{xy}^{11,11} & \Delta_{xy}^{11,12} & \dots & \Delta_{xy}^{11,t1} & \Delta_{xy}^{11,t2} \\ \Delta_{xy}^{12,11} & \Delta_{xy}^{12,12} & \dots & \Delta_{xy}^{12,t1} & \Delta_{xy}^{12,t2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \Delta_{xy}^{t1,11} & \Delta_{xy}^{t1,12} & \dots & \Delta_{xy}^{t1,t1} & \Delta_{xy}^{t1,t2} \\ \Delta_{xy}^{t2,11} & \Delta_{xy}^{t2,12} & \dots & \Delta_{xy}^{t2,t1} & \Delta_{xy}^{t2,t2} \end{pmatrix} \end{array},$$

$$\Delta_{xy,y} = \begin{array}{c} \\ p \\ p \\ \vdots \\ p \\ p \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 1 & \dots & 1 & 1 \end{array} \\ \begin{pmatrix} \Delta_{xy,y}^{11,11} & \Delta_{xy,y}^{11,12} & \dots & \Delta_{xy,y}^{11,t1} & \Delta_{xy,y}^{11,t2} \\ \Delta_{xy,y}^{12,11} & \Delta_{xy,y}^{12,12} & \dots & \Delta_{xy,y}^{12,t1} & \Delta_{xy,y}^{12,t2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \Delta_{xy,y}^{t1,11} & \Delta_{xy,y}^{t1,12} & \dots & \Delta_{xy,y}^{t1,t1} & \Delta_{xy,y}^{t1,t2} \\ \Delta_{xy,y}^{t2,11} & \Delta_{xy,y}^{t2,12} & \dots & \Delta_{xy,y}^{t2,t1} & \Delta_{xy,y}^{t2,t2} \end{pmatrix} \end{array},$$

$$\Delta_{y} = \begin{array}{c} \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 1 & \dots & 1 & 1 \end{array} \\ \begin{pmatrix} \Delta_{y}^{11,11} & \Delta_{y}^{11,12} & \dots & \Delta_{y}^{11,t1} & \Delta_{y}^{11,t2} \\ \Delta_{y}^{12,11} & \Delta_{y}^{12,12} & \dots & \Delta_{y}^{12,t1} & \Delta_{y}^{12,t2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \Delta_{y}^{t1,11} & \Delta_{y}^{t1,12} & \dots & \Delta_{y}^{t1,t1} & \Delta_{y}^{t1,t2} \\ \Delta_{y}^{t2,11} & \Delta_{y}^{t2,12} & \dots & \Delta_{y}^{t2,t1} & \Delta_{y}^{t2,t2} \end{pmatrix} \end{array},$$

$$\Delta_{xy,x} = \begin{array}{c} \\ p \\ p \\ \vdots \\ p \\ p \end{array} \begin{array}{c} p \\ \begin{pmatrix} \Delta_{xy,x}^{11} \\ \Delta_{xy,x}^{12} \\ \vdots \\ \Delta_{xy,x}^{t1} \\ \Delta_{xy,x}^{t2} \end{pmatrix} \end{array}, \quad \Delta_{y,x} = \begin{array}{c} \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{array} \begin{array}{c} p \\ \begin{pmatrix} \Delta_{y,x}^{11} \\ \Delta_{y,x}^{12} \\ \vdots \\ \Delta_{y,x}^{t1} \\ \Delta_{y,x}^{t2} \end{pmatrix} \end{array}.$$

$$\text{Let} \quad A = \begin{array}{c} \\ p \\ p \\ \vdots \\ p \\ p \end{array} \begin{array}{c} \begin{array}{cccccccccc} p & p & \dots & p & p & 1 & 1 & \dots & 1 & 1 \end{array} \\ \begin{pmatrix} I_p & 0 & \dots & 0 & 0 & \boldsymbol{\mu} & 0 & \dots & 0 & 0 \\ 0 & I_p & \dots & 0 & 0 & 0 & \boldsymbol{\mu} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & I_p & 0 & 0 & 0 & \dots & \boldsymbol{\mu} & 0 \\ 0 & 0 & \dots & 0 & I_p & 0 & 0 & \dots & 0 & \boldsymbol{\mu} \end{pmatrix} \end{array} \begin{array}{c} p \\ \mathrm{E}\cos(\omega_1^{\mathrm{T}} Y) I_p \\ \mathrm{E}\sin(\omega_1^{\mathrm{T}} Y) I_p \\ \vdots \\ \mathrm{E}\cos(\omega_t^{\mathrm{T}} Y) I_p \\ \mathrm{E}\sin(\omega_t^{\mathrm{T}} Y) I_p \end{array}.$$

Then, $\sqrt{n}\mathrm{vec}[\hat{\Theta} - \Theta + \boldsymbol{\mu}(Q - \hat{Q})^{\mathrm{T}} + (\boldsymbol{\mu} - \bar{\boldsymbol{x}})Q^{\mathrm{T}}] \sim N_{2pt}(\mathbf{0}, A\tau A^{\mathrm{T}})$.

Hence, $\sqrt{n}\text{vec}\{\Gamma_0^{\mathrm{T}}\Sigma^{-1/2}[\hat{\Theta}-\Theta+\boldsymbol{\mu}(Q-\hat{Q})^{\mathrm{T}}+(\boldsymbol{\mu}-\bar{\boldsymbol{x}})Q^{\mathrm{T}}]\Phi_0\} = (\Phi_0^{\mathrm{T}}\otimes\Gamma_0^{\mathrm{T}}\Sigma^{-1/2})\sqrt{n}\text{vec}[\hat{\Theta}-\Theta+\boldsymbol{\mu}(Q-\hat{Q})^{\mathrm{T}}+(\boldsymbol{\mu}-\bar{\boldsymbol{x}})Q^{\mathrm{T}}]$, which has normal distribution $N_{(2t-d)\times(p-d)}(\mathbf{0},\Omega=(\Phi_0^{\mathrm{T}}\otimes\Gamma_0^{\mathrm{T}}\Sigma^{-1/2})A\tau A^{\mathrm{T}}(\Phi_0\otimes\Sigma^{-1/2}\Gamma_0))$. Let $\Psi_0=\Gamma_0^{\mathrm{T}}\hat{\Psi}\Phi_0$, then $\hat{\Lambda}_d = n\text{trace}(\Psi_0\Psi_0^{\mathrm{T}}) = n\text{vec}(\Psi_0)^{\mathrm{T}}\text{vec}(\Psi_0)$. Because Ω is a positive definite matrix, there exists an orthogonal matrix P and diagonal matrix D such that $\Omega = P^{\mathrm{T}}DP$. Because $\sqrt{n}\text{vec}(\Psi_0)\sim N(\mathbf{0},\Omega)$, then $\sqrt{n}P\text{vec}(\Psi_0)\sim N(\mathbf{0},D)$. So $n\text{vec}(\Psi_0)^{\mathrm{T}}\text{vec}(\Psi_0)\sim\sum_{k=1}^{(p-d)(2t-d)}\lambda_k C_k$, where the C_ks are independent chi-square random variables with one degree of freedom and $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_{(p-d)(2t-d)}$ are eigenvalues of the covariance matrix Ω.

Remark: When $\mathbf{X}$ follows multivariate normal distribution, Proposition 2.2 will not result in a chi-square distribution. In fact, $\hat{\boldsymbol{\psi}}_{j1}=\frac{1}{n}\sum_{k=1}^{n}\mathbf{Z}_k\cos(\boldsymbol{\omega}_j^{\mathrm{T}}\mathbf{y}_k)$ is not a linear transformation of normal distribution. Hence, it is not a normal distribution. However, because the sample mean in different slices (Li 1991) is independent normal under assumption, we expect that the test statistic follows chi-square distribution. ■

Proof of Proposition 3.1**:** The proof is similar to the proof of Proportion 4.2 of Chiaromonte et al. (2002). First, fix the number of transformations $\{t_k\}_{k=1}^{K}$ and choose $\{\boldsymbol{\omega}_j^{(k)}\}_{j=1}^{t_k}$ for each level. For each level $k, j=1,\ldots,t_k$, define:

$$\hat{\boldsymbol{\psi}}_{j1}^{(k)}=\frac{1}{n_k}\sum_{l=1}^{n_k}\hat{\mathbf{z}}_l^{(k)}\cos(\boldsymbol{\omega}_j^{(k)T}\mathbf{y}_l^{(k)}),\quad \boldsymbol{\psi}_{j1}^{(k)}=\mathrm{E}[\mathbf{Z}\cos(\boldsymbol{\omega}_j^{(k)T}\mathbf{Y})],$$

$$\hat{\boldsymbol{\psi}}_{j2}^{(k)}=\frac{1}{n_k}\sum_{l=1}^{n_k}\hat{\mathbf{z}}_l^{(k)}\sin(\boldsymbol{\omega}_j^{(k)T}\mathbf{y}_l^{(k)}),\quad \boldsymbol{\psi}_{j2}^{(k)}=\mathrm{E}[\mathbf{Z}\sin(\boldsymbol{\omega}_j^{(k)T}\mathbf{Y})].$$

Let $\hat{\Psi}_k=(\hat{\boldsymbol{\psi}}_{11}^{(k)},\hat{\boldsymbol{\psi}}_{12}^{(k)},\ldots,\hat{\boldsymbol{\psi}}_{t1}^{(k)},\hat{\boldsymbol{\psi}}_{t2}^{(k)})$ and $\Psi_k=(\boldsymbol{\psi}_{11}^{(k)},\boldsymbol{\psi}_{12}^{(k)},\ldots,\boldsymbol{\psi}_{t1}^{(k)},\boldsymbol{\psi}_{t2}^{(k)})$. Let $\hat{f}_k=\sqrt{\frac{n_k}{n}}$, $\Psi^W=(f_1\Psi_1,\ldots,f_K\Psi_K)$ and $\hat{\Psi}^W=(\hat{f}_1\hat{\Psi}_1,\ldots,\hat{f}_K\hat{\Psi}_K)$. For each level $k, j=1,\ldots,t_k$, define:

$$\hat{\boldsymbol{\theta}}_{j1}^{(k)}=\frac{1}{n_k}\sum_{l=1}^{n_k}\mathbf{x}_l^{(k)}\cos(\boldsymbol{\omega}_j^{(k)T}\mathbf{y}_l^{(k)}),\quad \boldsymbol{\theta}_{j1}^{(k)}=\mathrm{E}[\mathbf{X}\cos(\boldsymbol{\omega}_j^{(k)T}\mathbf{Y})],$$

$$\hat{\boldsymbol{\theta}}_{j2}^{(k)}=\frac{1}{n_k}\sum_{l=1}^{n_k}\mathbf{x}_l^{(k)}\sin(\boldsymbol{\omega}_j^{(k)T}\mathbf{y}_l^{(k)}),\quad \boldsymbol{\theta}_{j2}^{(k)}=\mathrm{E}[\mathbf{X}\sin(\boldsymbol{\omega}_j^{(k)T}\mathbf{Y})].$$

Let $\hat{\Theta}_k=(\hat{\boldsymbol{\theta}}_{11}^{(k)},\hat{\boldsymbol{\theta}}_{12}^{(k)},\ldots,\hat{\boldsymbol{\theta}}_{t1}^{(k)},\hat{\boldsymbol{\theta}}_{t2}^{(k)})$ and $\Theta_k=(\boldsymbol{\theta}_{11}^{(k)},\boldsymbol{\theta}_{12}^{(k)},\ldots,\boldsymbol{\theta}_{t1}^{(k)},\boldsymbol{\theta}_{t2}^{(k)})$. Set $\hat{Q}_k=((1/n_k)\sum_{l=1}^{n_k}\cos(\boldsymbol{\omega}_1^{(k)T}\mathbf{y}_l^{(k)}),(1/n_k)\sum_{l=1}^{n_k}\sin(\boldsymbol{\omega}_1^{(k)T}\mathbf{y}_l^{(k)}),\ldots,(1/n_k)\sum_{l=1}^{n_k}\cos(\boldsymbol{\omega}_t^{(k)T}\mathbf{y}_l^{(k)}),(1/n_k)\sum_{l=1}^{n_k}\sin(\boldsymbol{\omega}_t^{(k)T}\mathbf{y}_l^{(k)}))^{\mathrm{T}}$, and $Q_k=(\mathrm{E}(\cos\boldsymbol{\omega}_1^{(k)T}\mathbf{Y}),\mathrm{E}(\sin\boldsymbol{\omega}_1^{(k)T}\mathbf{Y}),\ldots,\mathrm{E}(\cos\boldsymbol{\omega}_t^{(k)T}\mathbf{Y}),\mathrm{E}(\sin\boldsymbol{\omega}_t^{(k)T}\mathbf{Y}))^{\mathrm{T}}$. Then by SVD, $\Psi^W=\boldsymbol{\Gamma}^{\mathrm{T}}\left[\begin{smallmatrix}\boldsymbol{D}&0\\0&0\end{smallmatrix}\right]\boldsymbol{\Phi}$, where $\boldsymbol{\Gamma}$ and $\boldsymbol{\Phi}$ are respective $p\times p$ and $2\sum t_k\times 2\sum t_k$ orthogonal matrices and $\boldsymbol{D}$ is a $d\times d$ diagonal matrix of positive values. Let $\boldsymbol{\Gamma}^{\mathrm{T}}=(\boldsymbol{\Gamma}_1,\boldsymbol{\Gamma}_0)$ and $\boldsymbol{\Phi}^{\mathrm{T}}=(\boldsymbol{\Phi}_1,\boldsymbol{\Phi}_0)$, where $\boldsymbol{\Gamma}_0$ is $p\times(p-d)$ and $\boldsymbol{\Phi}_0$ is $2\sum t_k\times(2\sum t_k-d)$. Thus $\hat{\Lambda}_d^W=n\times\text{trace}[(\boldsymbol{\Gamma}_0^{\mathrm{T}}\hat{\Psi}^W\boldsymbol{\Phi}_0)(\boldsymbol{\Gamma}_0^{\mathrm{T}}\hat{\Psi}^W\boldsymbol{\Phi}_0)^{\mathrm{T}}]=n\ \text{vec}(\boldsymbol{\Gamma}_0^{\mathrm{T}}\hat{\Psi}^W\boldsymbol{\Phi}_0)^{\mathrm{T}}\ \text{vec}(\boldsymbol{\Gamma}_0^{\mathrm{T}}\hat{\Psi}^W\boldsymbol{\Phi}_0)$. Partition $\boldsymbol{\Phi}_0=(\boldsymbol{\Phi}_{01}^{\mathrm{T}},\ldots,\boldsymbol{\Phi}_{0K}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\Phi}_{0k}$ has dimension $2t_k\times(2\sum t_k-d)$. Then $\sqrt{n}\boldsymbol{\Gamma}_0^{\mathrm{T}}\hat{\Psi}^W\boldsymbol{\Phi}_0=\sqrt{n}\boldsymbol{\Gamma}_0^{\mathrm{T}}(\sum_{k=1}^{K}\hat{f}_k\hat{\Psi}_k\boldsymbol{\Phi}_{0k})=\sum_{k=1}^{K}\sqrt{n_k}\boldsymbol{\Gamma}_0^{\mathrm{T}}\hat{\Psi}_k\boldsymbol{\Phi}_{0k}$. As $\boldsymbol{\Gamma}_0^{\mathrm{T}}\Psi^W\boldsymbol{\Phi}_0=0$, that is $\boldsymbol{\Gamma}_0^{\mathrm{T}}\Psi^W\boldsymbol{\Phi}_0=\boldsymbol{\Gamma}_0^{\mathrm{T}}\sum_{k=1}^{K}f_k\Psi_k\boldsymbol{\Phi}_{0k}=\boldsymbol{\Gamma}_0^{\mathrm{T}}\sum_{k=1}^{K}f_k\Sigma^{-1/2}(\Theta_k-\boldsymbol{\mu}Q_k)\boldsymbol{\Phi}_{0k}=0$.

Define $\hat{A}=\hat{\Sigma}^{-1/2}\Sigma^{1/2}$, then

$$\sqrt{n}\Gamma_0^{\mathrm{T}}\hat{\Psi}^W\boldsymbol{\Phi}_0=\sqrt{n}\Gamma_0^{\mathrm{T}}\hat{\Sigma}^{-1/2}\sum_{k=1}^{K}\hat{f}_k(\hat{\Theta}_k-\bar{\boldsymbol{x}}^{(k)}\hat{Q}_k^{\mathrm{T}})\boldsymbol{\Phi}_{0k}$$

$$= \sqrt{n}\Gamma_0^{\mathrm{T}}\hat{A}\Sigma^{-1/2}\sum_{k=1}^{K}\hat{f}_k(\hat{\Theta}_k - \bar{\boldsymbol{x}}^{(k)}\hat{Q}_k^{\mathrm{T}})\boldsymbol{\Phi}_{0k}$$

$$= \sqrt{n}\Gamma_0^{\mathrm{T}}\hat{A}\Sigma^{-1/2}\sum_{k=1}^{K}\frac{\hat{f}_k}{f_k}f_k(\hat{\Theta}_k - \bar{\boldsymbol{x}}^{(k)}\hat{Q}_k^{\mathrm{T}})\boldsymbol{\Phi}_{0k}$$

$$= \sqrt{n}\Gamma_0^{\mathrm{T}}(\hat{A} - I + I)\Sigma^{-1/2}\sum_{k=1}^{K}\left(\frac{\hat{f}_k}{f_k} - 1 + 1\right)f_k(\hat{\Theta}_k - \bar{\boldsymbol{x}}^{(k)}\hat{Q}_k^{\mathrm{T}})\boldsymbol{\Phi}_{0k}$$

$$= \sqrt{n}\Gamma_0^{\mathrm{T}}(\hat{A} - I)\Sigma^{-1/2}\sum_{k=1}^{K}f_k(\hat{\Theta}_k - \bar{\boldsymbol{x}}^{(k)}\hat{Q}_k^{\mathrm{T}})\boldsymbol{\Phi}_{0k}$$

$$+ \sqrt{n}\Gamma_0^{\mathrm{T}}\Sigma^{-1/2}\sum_{k=1}^{K}\hat{f}_k(\hat{\Theta}_k - \bar{\boldsymbol{x}}^{(k)}\hat{Q}_k^{\mathrm{T}})\boldsymbol{\Phi}_{0k} + O_p\left(\frac{1}{n}\right).$$

So $\hat{\Theta}_k - \bar{\boldsymbol{x}}^{(k)}\hat{Q}_k^{\mathrm{T}} = \hat{\Theta}_k - \Theta_k + \Theta_k - \boldsymbol{\mu}Q_k^{\mathrm{T}} + \boldsymbol{\mu}(Q_k^{\mathrm{T}} - \hat{Q}_k^{\mathrm{T}}) + (\boldsymbol{\mu} - \bar{\boldsymbol{x}}^{(k)})\hat{Q}_k^{\mathrm{T}}$. Then we use $(\hat{A} - I)\Sigma^{-1/2}(\hat{\Theta}_k - \Theta_k) = O_p(1/n_k)$, $(\hat{A} - I)\Sigma^{-1/2}\boldsymbol{\mu}(Q_k^{\mathrm{T}} - \hat{Q}_k^{\mathrm{T}}) = O_p(1/n_k)$, $(\hat{A} - I)\Sigma^{-1/2}(\boldsymbol{\mu} - \bar{\boldsymbol{x}}_{(k)})\hat{Q}_k^{\mathrm{T}} = O_p(1/n)$ and $\sum_{k=1}^{K}f_k\Sigma^{-1/2}(\Theta_k - \boldsymbol{\mu}Q_k)\boldsymbol{\Phi}_{0k} = 0$.

$$\sqrt{n}\Gamma_0^{\mathrm{T}}\hat{\Psi}^W\boldsymbol{\Phi}_0 = \sqrt{n}\Gamma_0^{\mathrm{T}}\Sigma^{-1/2}\sum_{k=1}^{K}\hat{f}_k(\hat{\Theta}_k - \Theta_k + \boldsymbol{\mu}(Q_k^{\mathrm{T}} - \hat{Q}_k^{\mathrm{T}}) + (\boldsymbol{\mu} - \bar{\boldsymbol{x}}^{(k)})\hat{Q}_k^{\mathrm{T}})\boldsymbol{\Phi}_{0k} + O_p\left(\frac{1}{n}\right).$$

$$= \Gamma_0^{\mathrm{T}}\Sigma^{-1/2}\sum_{k=1}^{K}\sqrt{n_k}(\hat{\Theta}_k - \Theta_k + \boldsymbol{\mu}(Q_k^{\mathrm{T}} - \hat{Q}_k^{\mathrm{T}}) + (\boldsymbol{\mu} - \bar{\boldsymbol{x}}^{(k)})\hat{Q}_k^{\mathrm{T}})\boldsymbol{\Phi}_{0k} + O_p\left(\frac{1}{n}\right).$$

Let $\Omega_k = (\boldsymbol{\Phi}_{0k}^{\mathrm{T}} \otimes \boldsymbol{\Gamma}_0^{\mathrm{T}}\Sigma^{-1/2})A_k\tau_kA_k^{\mathrm{T}}(\boldsymbol{\Phi}_{0k} \otimes \Sigma^{-1/2}\boldsymbol{\Gamma}_0)$ and τ_k are defined in Proposition 2.2. Then $\sqrt{n}\boldsymbol{\Gamma}_0^{\mathrm{T}}\hat{\Psi}^W\boldsymbol{\Phi}_0 \sim N(\mathbf{0}, \sum\Omega_k)$. Furthermore, the rank for $\sum\Omega_k$ is $(p-d)(2\sum t_k - Kd)$. So $\hat{\Lambda}_d^W \sim \sum_{i=1}^{(p-d)(2\sum t_k - Kd)}\lambda_iC_i$, where the C_is are independent chi-square random variables, each with one degree of freedom, and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{(p-d)(2\sum t_k - Kd)}$ are eigenvalues of the covariance matrix $\Omega^W = \sum\Omega_k$. ■