Statistica Sinica Preprint No: SS-2020-0312	
Title	A Minimum Discrepancy Approach With Fourier Transform in Sufficient Dimension Reduction
Manuscript ID	SS-2020-0312
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0312
Complete List of Authors	Jiaying Weng and Xiangrong Yin
Corresponding Author	Jiaying Weng
E-mail	jweng@bentley.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Statistica Sinica

A minimum discrepancy approach with Fourier transform in sufficient dimension reduction

Jiaying Weng and Xiangrong Yin

Bentley University and University of Kentucky

Abstract:

We propose an optimal family of estimators in sufficient dimension reduction using a Fourier transform based on a quadratic discrepancy function. Our proposed approach has advantages over existing methods by avoiding the slicing scheme of a response variable and easily handling multivariate response. We further develop four sub-optimal estimators: the degenerated and special estimators for computational efficiency and simplicity, and the robust and its degenerated estimators for a less restrictive condition for estimation and inference. Marginal and conditional hypothesis tests for predictors and dimensions are also obtained. Simulation studies and real data analysis are carried out to illustrate the efficacy of our proposed methods.

Key words and phrases: Fourier transform; Minimum discrepancy; Sufficient dimension reduction.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

1. Introduction

With the recent development in data collection and storage techniques, vast data with huge volume and high dimension are available for researchers to build an economic model and create an advanced visualization tool. It is easier to achieve these goals if we can obtain a low-dimensional function of the predictor associated with the response variable. This paper focuses on sufficient dimension reduction (SDR; Li 1991; Cook 1996), a model-free approach. It preserves complete regression information, making it attractive to researchers to build a parsimonious model.

SDR considers a regression of $q\times 1$ response $\mathbf{Y}$ given a $p\times 1$ predictor $\mathbf{X}$. It aims to find a dimension reduction matrix $\beta \in \mathbb{R}^{p\times d}(d \leq p)$ such that the reduced variables $\beta^T\mathbf{X}$ retain complete regression information. Matrix β may not be identifiable, but the space spanned by the columns of β, known as the dimension reduction subspace and denoted as Span(β), is identifiable. To achieve the uniqueness and minimum of the subspace, we study the central subspace $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$, which is the intersection of all dimension reduction subspaces if the intersection itself is a dimension reduction subspace. Under mild conditions, (Cook, 1996; Yin, Li, and Cook, 2008), the central subspace exists and is unique.

Sliced inverse regression (SIR; Li 1991) and sliced average variance es-

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

timation (SAVE; Cook and Weisberg 1991) are the first two methods proposed for SDR. The main idea of SIR or SAVE is to identify a dimension reduction kernel matrix, using its eigenvectors and structural dimension, say, d, to estimate the central subspace. Motivated by these two methods, other various dimension reduction kernel matrices have been proposed (Cook, 1998a; Chiaromonte et al., 2002; Ye and Weiss, 2003; Li and Wang, 2007a; Wang and Xia, 2008). Regarding the statistical inference, Cook (2004) studied hypothesis tests of the effectiveness of selected predictors in regression. Along with this idea, Cook and Ni (2005) introduced a novel optimal estimator via minimum discrepancy function (IRE). Since then, numerous IR-based methods have been employed to SDR problems (Cook and Ni, 2005; Ni and Cook, 2007; Cook and Zhang, 2014; Qian et al., 2019).

Extending SDR methods to multivariate response is a challenge because it encounters the curse of dimensionality (Cook and Setodji, 2003; Saracco, 2005; Yin and Bura, 2006). The merit of Fourier transform has been brought to researchers' attention, especially in multivariate responses (Zhu and Zeng, 2006; Zhu et al., 2010). Weng and Yin (2018) deliberated the Fourier transform (FT) in various scenarios, along with the partial SDR (Chiaromonte et al., 2002; Li et al., 2005) and sequential SDR (Yin and Hilafu, 2015).

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

In this paper, we propose a novel family of optimal estimators based on the optimization of the quadratic function with Fourier transform approach. Our proposal differs from past methods in at least three aspects. First, the minimum discrepancy with Fourier transform has not been discussed in the literature before. Our work fills this gap. We explore not only an optimal estimator in SDR problems but also the other four sub-optimal estimators "degenerated", "special", "robust", and "degenerated robust". The degenerated and special estimators are more computational efficient after simplifying the calculation of the precision covariance matrix. Besides, the robust and its degenerated estimator only require second moments of the predictor in the procedures of estimation and inference. Second, we provide an entire package about the minimum discrepancy with Fourier transform approach in SDR by investigating the conditional and marginal hypothesis tests to identify structural dimensions and the significance of predictors. Last but not least, the singularity of the limiting covariance matrix is discussed when repeated information occurs. And, we compare two approaches to estimate the precision matrix and, at the same time, maintain the asymptotic efficiency and chi-squared test statistics.

The rest of the article is organized as follows. We review the Fourier transform approach and construct a minimal discrepancy approach with

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Fourier transform in Section 2. The other four sub-optimal estimators are described in Section 3. In Section 4, simulation studies, and real data analysis are presented to support our theoretical analysis. We conclude with a discussion in Section 5. All proofs, unless, otherwise stated, and addition simulation results are in the supplementary file.

2. Optimal Estimator with Fourier Transform

In this section, we introduce a new optimal family of minimum discrepancy with Fourier transform approach in SDR and discuss related properties.

2.1 Fourier Transform via Kernel Matrix

Suppose matrix $\beta \in \mathbb{R}^{p\times d}$ is a basis of the central subspace, where d is the structural dimension. Define $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}$ as $\mathbf{Y}$ is statistically independent of $\mathbf{X}$. Our goal is to estimate β such that $\beta^T\mathbf{X}$ fully describe the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$. The following properties about the central subspace $\mathcal{S}_{\mathbf{Y}|\mathbf{X}} = \text{Span}(\beta)$ are the keys to ensure above goal feasible (Cook, 1998b; Cook and Zhang, 2014): $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\beta^T\mathbf{X} \iff \mathbf{X}|(\mathbf{Y}, \beta^T\mathbf{X}) \sim \mathbf{X}|\beta^T\mathbf{X} \iff \mathbf{Y}|\mathbf{X} \sim \mathbf{Y}|\beta^T\mathbf{X}$. These properties show that using $\beta^T\mathbf{X}$ as reduced new variables to fit a regression model is sufficient and efficient.

Three approaches, inverse, forward and joint, have been developed for

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

SDR to solve β. The inverse approach uses the regression of $\mathbf{X}$ on $\mathbf{Y}$. One way is to estimate $\mathrm{E}(\mathbf{X}|\mathbf{Y})$ and to categorize $\mathbf{Y}$ by slicing, for example, SIR (Li, 1991). In this paper, we focus on the inverse approach through Fourier transform (Zhu et al., 2010; Weng and Yin, 2018).

Let $\mathbf{X}_0 = \mathbf{X} - \mathrm{E}(\mathbf{X})$ and Σ be the covariance matrix of $\mathbf{X}$. The Fourier transform of $\mathrm{E}(\mathbf{X}_0|\mathbf{Y})f(\mathbf{Y})$ giving a frequency $\boldsymbol{\omega} \in \mathbb{R}^q$ is

$$\phi(\boldsymbol{\omega}) = \int e^{i\boldsymbol{\omega}^T\mathbf{Y}}\mathrm{E}(\mathbf{X}_0|\mathbf{Y})f(\mathbf{Y})d\mathbf{Y} = \mathrm{E}(e^{i\boldsymbol{\omega}^T\mathbf{Y}}\mathbf{X}) - \mathrm{E}(e^{i\boldsymbol{\omega}^T\mathbf{Y}})\mathrm{E}(\mathbf{X}),$$

where $f(\mathbf{Y})$ is the marginal density of $\mathbf{Y}$. Under the well-known linearity condition: $\mathrm{E}(\mathbf{X}|\beta^T\mathbf{X})$ is a linear function of $\beta^T\mathbf{X}$, both the real and imaginary parts of $\Sigma^{-1}\phi(\boldsymbol{\omega})$ are in the central subspace. It can be shown that if $\mathbf{X}$ has an elliptically contoured distribution, then the linearity condition is satisfided (Li, 2018). The linearity condition ensures $\mathrm{Span}\{\Sigma^{-1}\mathrm{E}(e^{i\boldsymbol{\omega}^T\mathbf{Y}}\mathbf{X}_0), \boldsymbol{\omega} \in \mathbb{R}^q\} \subseteq \mathcal{S}_{\mathbf{Y}|\mathbf{X}}$. To facilitate the discussion, we use the commonly assumed coverage condition (Cook and Ni, 2005; Li and Wang, 2007b): $\mathrm{Span}\{\Sigma^{-1}\mathrm{E}(e^{i\boldsymbol{\omega}^T\mathbf{Y}}\mathbf{X}_0), \boldsymbol{\omega} \in \mathbb{R}^q\} = \mathcal{S}_{\mathbf{Y}|\mathbf{X}}$, which enforces the equality of two subspaces.

The $\phi(\boldsymbol{\omega})$ is superior to the existing inverse regression approaches in following three folds: (1) estimating $\mathrm{E}(e^{i\boldsymbol{\omega}^T\mathbf{Y}}\mathbf{X})$ avoids a slicing scheme of $\mathbf{Y}$, (2) the Fourier transform can easily deal with a multivariate response because $\boldsymbol{\omega}^T\mathbf{Y}$ transforms a multivariate response to a scalar, (3) one $\boldsymbol{\omega}$ pro-

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

vides two vectors in the central subspace: the real part $\Sigma^{-1}\mathrm{E}[\cos(\boldsymbol{\omega}^T\mathbf{Y})\mathbf{X}_0]$ and the imaginary part $\Sigma^{-1}\mathrm{E}[\sin(\boldsymbol{\omega}^T\mathbf{Y})\mathbf{X}_0]$.

Zhu et al. (2010) and Weng and Yin (2018) constructed a following kernel matrix via $\phi(\boldsymbol{\omega})$ to estimate a basis of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$,

$$\left\{\Sigma^{-1}\mathrm{E}[\cos(\boldsymbol{\omega}^T\mathbf{Y})\mathbf{X}_0], \Sigma^{-1}\mathrm{E}[\sin(\boldsymbol{\omega}^T\mathbf{Y})\mathbf{X}_0],\ \boldsymbol{\omega}\in\mathbb{R}^q\right\}.$$

The matrix that consists of the eigenvectors corresponding to its largest d eigenvalues is one estimate of the basis of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$, denoted by FT (Weng and Yin, 2018).

2.2 The choice of ω

We randomly generate $\boldsymbol{\omega}$ from $N(\mathbf{0}, \frac{s\pi^2}{\mathrm{E}(\mathbf{Y}^T\mathbf{Y})}I)$ satisfying $P(|\boldsymbol{\omega}^T\mathbf{Y}| > \pi) \leq s$, meaning that the probability of obtaining repeated information is less than s. Random generation makes it possible to recover all potential directions in the central subspace. Similar ideas have been implemented to deal with a multivariate response in sufficient dimension reduction. For example, Li et al. (2008) introduced a projective resampling method based on t^TY, where t is generated uniformly from a unit sphere.

Based on the simulation study presented in Zhu et al. (2010), estimator has a stable and satisfactory performance when $0.02 < s < 0.30$. The effect of s values is empirical insensitive on the estimation. We choose a moderate

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

value, 0.1, such that the probability of providing additional information in recovering the central subspaces is greater than 0.9. Other than the s value, the choice of $\boldsymbol{\omega}$ is also determined by the magnitude of $\mathbf{Y}^T\mathbf{Y}$. If the variance of $\boldsymbol{\omega}$ is too small, the subspace $\text{Span}\{\Sigma^{-1}\text{E}(e^{i\boldsymbol{\omega}^T\mathbf{Y}}\mathbf{X}_0) : \boldsymbol{\omega} \in \mathbb{R}^q\}$ is close to the null space and will miss some directions in the central space. To avoid extreme response values, we suggest that if the ratio of the sample mean of $\mathbf{Y}^T\mathbf{Y}$ to its sample median is greater than a threshold value, say 100, and then we prefer the sample median; otherwise, we use the mean. For instance, the models $Y = \exp(\mathbf{X}^T\beta) + \epsilon$ and $Y = \frac{1}{\mathbf{X}^T\beta} + \epsilon$ have more extreme Y values or outliers, so the median is better.

Here, $\boldsymbol{\omega}$ is a tuning parameter in Fourier transformation. Weng and Yin (2018) showed that a finite and large enough number of $\boldsymbol{\omega}$s is sufficient to recover the central subspace and suggested that moderate m suffices to have excellent performance. Our limited simulation also supports that the proposed methods are insensitive to the choice of $\boldsymbol{\omega}$ when the number is sufficient.

2.3 A General Framework: Fourier Transform via MDF

Previously, we reviewed the Fourier transform approach based on a kernel matrix applying singular value decomposition. However, this estimator can

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

be obtained as a special case via the minimum discrepancy function (MDF, Cook and Ni 2005; Cook and Zhang 2014). MDF is the minimization of a quadratic discrepancy function characterizing the difference between the sample and population values. Cook and Ni (2005) proposed the asymptotic optimal inverse regression estimator (IRE), however, it relies on a slicing scheme of $\mathbf{Y}$. Cook and Zhang (2014) reduced the complexity of choosing a good slicing scheme by introducing fusing methods (FIRE and DIRE). These methods still suffer the curse of dimension when the response is multivariate. Now we follow the framework of Cook and Ni (2005) and Cook and Zhang (2014) to generalize the Fourier transform via MDF.

Suppose the size of $\boldsymbol{\omega}$s is m, $\{\boldsymbol{\omega}_j\}_{j=1}^m$ is generated as in the previous section. Let $\boldsymbol{\xi}_j = \Sigma^{-1}[\mathrm{E}(e^{i\boldsymbol{\omega}_j^T\mathbf{Y}}\mathbf{X}) - \mathrm{E}(e^{i\boldsymbol{\omega}_j^T\mathbf{Y}})\mathrm{E}(\mathbf{X})] \in \mathbb{C}^p$. Let indexes R and I represent the real and imaginary parts of a complex vector, and $\xi = \left(\boldsymbol{\xi}_j^R, \boldsymbol{\xi}_j^I\right)_{j=1}^m$ denotes a matrix combining each $\boldsymbol{\xi}_j^R$ and $\boldsymbol{\xi}_j^I$. The working meta-parameter is defined as $\mathcal{S}_\xi = \sum_{j=1}^m \mathrm{Span}(\boldsymbol{\xi}_j^R, \boldsymbol{\xi}_j^I)$. Note that β is a basis of $\mathcal{S}_\xi$ under the coverage condition, there exists a vector $\boldsymbol{\gamma}_j \in \mathbb{C}^p$ such that $\boldsymbol{\xi}_j = \beta\boldsymbol{\gamma}_j$ for each j. Let $\nu = \left(\boldsymbol{\gamma}_j^R, \boldsymbol{\gamma}_j^I\right)_{j=1}^m$ be the matrix combining each $\boldsymbol{\gamma}_j^R$ and $\boldsymbol{\gamma}_j^I$. Thus, $\xi = \beta\nu$.

The corresponding sample values can be achieved using the sample

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

mean. Suppose $\{\mathbf{y}_i, \mathbf{x}_i\}$, $i = 1, ..., n$ are i.i.d. samples of $(\mathbf{Y}, \mathbf{X})$, and $\bar{\mathbf{x}}$ and

$$\hat{\boldsymbol{\xi}}_j = \hat{\Sigma}^{-1} \left(\frac{1}{n} \sum_{k=1}^{n} e^{i\boldsymbol{\omega}_j^T \mathbf{y}_k} \mathbf{x}_k - \frac{1}{n} \sum_{k=1}^{n} e^{i\boldsymbol{\omega}_j^T \mathbf{y}_k} \bar{\mathbf{x}} \right)$$

denote the sample mean of $\mathbf{X}$ and the sample estimate of $\boldsymbol{\xi}_j$, respectively. Then, $\hat{\xi} = \left(\hat{\boldsymbol{\xi}}_j^R, \hat{\boldsymbol{\xi}}_j^I \right)_{j=1}^{m}$ is a sample estimate of ξ, and we define the Fourier transform quadratic discrepancy function (QDF) of $B \in \mathbb{R}^{p \times d}, C \in \mathbb{R}^{d \times 2m}$ given the inner product matrix V as:

$$F_d(B, C; V) = [\text{vec}(\hat{\xi}) - \text{vec}(BC)]^T V [\text{vec}(\hat{\xi}) - \text{vec}(BC)]. \tag{2.1}$$

The minimization $(\hat{\beta}, \hat{\nu}) = \arg\min F_d(B, C; V)$ are the estimate $\hat{\beta}$, an orthogonal basis of the central subspace, and the coordinates $\hat{\nu}$ of $\hat{\xi}$ relative to the basis $\hat{\beta}$. We will discuss the choice of the inner product matrix V which may lead to different estimators, and show the respective asymptotic properties of the estimates.

2.4 Properties of FT-IRE

Note that $\hat{\xi}$ is a matrix, so the vectorization of $\hat{\xi}$, $\text{vec}(\hat{\xi})$, is employed in deriving the asymptotic theorem. Define $\mathbf{Z} = \Sigma^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$. Although we start with $\mathbf{X}$, in proving or state the final result we may use $\mathbf{Z}$ for simple presentations. As a result, the following theorem is fundamental and important.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Theorem 1. *Assume that* $\{\mathbf{y}_k, \mathbf{x}_k\}$, $k = 1, ..., n$, *are random samples of* $(\mathbf{Y}, \mathbf{X})$ *with finite fourth moments. Let* $\epsilon_j = e^{i\boldsymbol{\omega}_j^T \mathbf{Y}} - \mathrm{E}e^{i\boldsymbol{\omega}_j^T \mathbf{Y}} - \mathbf{Z}^T \mathrm{E}(e^{i\boldsymbol{\omega}_j^T \mathbf{Y}} \mathbf{Z})$ *be the population residual from ordinary least squares fit of* $e^{i\boldsymbol{\omega}_j^T \mathbf{Y}}$ *on* $\mathbf{Z}$, *and* $\boldsymbol{\epsilon} = \left(\epsilon_1^R, \epsilon_1^I, ..., \epsilon_m^R, \epsilon_m^I\right)^T$ *consists of its real and imaginary. Then*

$$\sqrt{n}[\mathrm{vec}(\hat{\xi}) - \mathrm{vec}(\beta\nu)] \xrightarrow{D} N(0, \Gamma),$$

where $\Gamma = \mathrm{Cov}\{\mathrm{vec}[\Sigma^{-1/2} \mathbf{Z} \boldsymbol{\epsilon}^T]\} \in \mathbb{R}^{2pm \times 2pm}$.

When minimizing QDF using the information matrix $V = \Gamma^{-1}$, the estimate $\hat{\beta}$ is called a Fourier Transform Inverse Regression Estimator (FT-IRE). To prove the asymptotic distribution of MDF, there are two basic assumptions, 1) identifiability of the parameters $\theta = (\mathrm{vec}(\beta)^T, \mathrm{vec}(\nu)^T)^T$ and 2) nonsingularity of the information matrix corresponding to $g(\theta) = \mathrm{vec}(\beta\nu)$. Our models suffer redundant parameters, overparameterized models, so these two assumptions are violated. Based on Shapiro (1986), we still can derive an asymptotic chi-squared distribution of the MDF test statistics and asymptotic distribution of the MDF estimators if Γ is invertible. The limiting covariance matrix in Theorem 1 can be written as,

$$\begin{aligned}
& \mathrm{Cov}\{\mathrm{vec}[\Sigma^{-1}(\mathbf{X} - \boldsymbol{\mu})\boldsymbol{\epsilon}^T]\} && = \mathrm{Cov}[\mathrm{vec}(\Sigma^{-1/2} \mathbf{Z} \boldsymbol{\epsilon}^T)] \\
= \; & \mathrm{Cov}[(I_{2m} \otimes \Sigma^{-1/2}) \mathrm{vec}(\mathbf{Z}\boldsymbol{\epsilon}^T)] && = (I_{2m} \otimes \Sigma^{-1/2}) \mathrm{Cov}[\mathrm{vec}(\mathbf{Z}\boldsymbol{\epsilon}^T)](I_{2m} \otimes \Sigma^{-1/2}) \\
= \; & (I_{2m} \otimes \Sigma^{-1/2}) \mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T \otimes \mathbf{Z}\mathbf{Z}^T)(I_{2m} \otimes \Sigma^{-1/2}) &&
\end{aligned}$$

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Note that, $\boldsymbol{\epsilon}$ is uncorrelated with $\mathbf{Z}$, that is $\mathrm{Cov}(\boldsymbol{\epsilon}, \mathbf{Z}) = 0$ and $\mathrm{E}(\boldsymbol{\epsilon}) = 0$. We further assume that $\boldsymbol{\epsilon}$ is independent of $\mathbf{Z}$, that is a common assumption in the regression problem. Hence, $\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T \otimes \mathbf{Z}\mathbf{Z}^T) = \mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T) \otimes I_p$ and $\Gamma = (I_{2m} \otimes \Sigma^{-1/2})[\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T) \otimes I_p](I_{2m} \otimes \Sigma^{-1/2}) = \mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T) \otimes \Sigma^{-1}$, which leads to $V = \mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)^{-1} \otimes \Sigma$. The singularity of Γ depends on the covariance matrix of $\boldsymbol{\epsilon}$. Each component of $\boldsymbol{\epsilon}$ can be regarded as the real and imaginary parts of the population residual of $e^{i\boldsymbol{\omega}^T\mathbf{Y}}$ regression on $\mathbf{Z}$. The way we generate $\boldsymbol{\omega}$ guarantees that each $\boldsymbol{\omega}^T\mathbf{Y}$ provides different random variable within one period $[-\pi, \pi]$, so $e^{i\boldsymbol{\omega}^T\mathbf{Y}}$ should be different from each other. However, if there are some duplicated information, ideally we can omit them from the matrix. Hence, under mild conditions, we can conclude that Γ is nonsingular.

Theorem 2. *Assume that* $\{\mathbf{y}_k, \mathbf{x}_k\}$, $k = 1, ..., n$, *are random samples of* $(\mathbf{Y}, \mathbf{X})$ *with finite fourth moments. Let* $(\hat{\beta}, \hat{\nu}) = \arg\min_{B,C} F_d(B, C; \hat{\Gamma}^{-1})$, *where* $\hat{\Gamma}$ *is a consistent estimate of* Γ. *Then the following results hold:*

1. $\mathrm{vec}(\hat{\beta}\hat{\nu})$ *is asymptotically efficient, and* $\sqrt{n}[\mathrm{vec}(\hat{\beta}\hat{\nu}) - \mathrm{vec}(\beta\nu)]$ *is an asymptotic normal with mean 0 and covariance matrix* $\Delta(\Delta^T V \Delta)^- \Delta^T$, *where* $V = \Gamma^{-1}$ *and* $\Delta = (v^T \otimes I_p, I_{2m} \otimes \beta)$ *with* $2mp \times d(p + 2m)$ *dimensions.*

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

2. *$n\hat{F}_d$ has an asymptotic chi-squared distribution with degrees of freedom* $(p-d)(2m-d)$.

3. $\text{Span}(\hat{\beta})$ *is a consistent estimator of* $\mathcal{S}_\xi$.

Part 2 of Theorem 2 can be used to estimate dimensions using the sequential tests. Let $T_k = n\hat{F}_k$ be a test statistic for $H_0 : d = k$ vs. $H_a : d > k$, where k is an integer from 0 to p. Note that T_k follows the asymptotic chi-squared distribution with the degrees of freedom $(p-k)(2m-k)$. The first time when we fail to reject the null hypothesis, e.g., $H_0 : d = d_0$, we can conclude that the dimension should be d_0. Also, Theorem 2 requires the covariance matrix Γ to be nonsingular. If it is not, we use two approaches.

1). We employ the generalized inverse matrix of Γ. That is $\Gamma^- = UD^-U^T$, where the columns of U are the eigenvectors of Γ and the diagonal matrix D has diagonal elements corresponding to the eigenvalues of Γ. The diagonal elements of D^- are reciprocal of eigenvalues if they are not 0s, otherwise are 0s. The eigenvalues 0s indicate that the corresponding columns of ξ are linearly correlated, equivalently to delete those correlated columns and replace with a smaller number of m.

2). We apply QL decomposition of $\Gamma = QL$ (assuming the sparsity of Γ):

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

the product of an unitary matrix Q (that is $QQ^T = Q^TQ = I$) and a lower triangular matrix L. So $\Gamma^{-1} = L^{-1}Q^T$. When m is small (says 4 or 6 when $d = 1, 2$), using the QL decomposition of soft-thresholding covariance Γ (Bickel and Levina, 2008; Rothman et al., 2009) can provide more accurate estimation. The detailed algorithm is in the supplementary file.

2.5 Predictor Hypothesis Tests of FT-IRE

We now investigate the conditional independence hypothesis test for FT-IRE. Suppose $\mathcal{H}$ is a user-specified subspace for predictors with r dimensions. Let $P_{\mathcal{H}}$ be the orthogonal projection onto $\mathcal{H}$ in the usual inner product, and $Q_{\mathcal{H}} = I - P_{\mathcal{H}}$ be the orthogonal projection onto the orthogonal complement of $\mathcal{H}$. A conditional independence hypothesis means that $H_0 : \mathbf{Y} \perp\!\!\!\perp P_{\mathcal{H}}\mathbf{X}|Q_{\mathcal{H}}\mathbf{X}$. Following Cook (2004) and Cook and Zhang (2014), we have

$$\mathbf{Y} \perp\!\!\!\perp P_{\mathcal{H}}\mathbf{X}|Q_{\mathcal{H}}\mathbf{X} \Leftrightarrow P_{\mathcal{H}}\mathcal{S}_{\mathbf{Y}|\mathbf{X}} = \mathcal{O}_p \Leftrightarrow P_{\mathcal{H}}\mathcal{S}_{\xi} = \mathcal{O}_p \Leftrightarrow H^T\xi = \mathbf{0},$$

where $\mathcal{O}_p$ is the origin in $\mathbb{R}^p$ and $H \in \mathbb{R}^{p\times r}$ is a basis for $\mathcal{H}$. Note that if $r + \dim(\mathcal{S}_{\mathbf{Y}|\mathbf{X}}) > p$, $P_{\mathcal{H}}\mathcal{S}_{\mathbf{Y}|\mathbf{X}} = \mathcal{O}_p$ can never be true. Hence, the following discussion only considers $r \leq p - \dim(\mathcal{S}_{\mathbf{Y}|\mathbf{X}})$. We discuss three hypothesis tests, their test statistics, and asymptotic distributions based on the

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

framework of Cook and Ni (2005).

1. Marginal predictor hypotheses, $H_0 : H^T\xi = 0$ vs. $H_a : H^T\xi \neq 0$. It only tests on predictors without requiring the specific value d. Theorem 1 indicates that $\sqrt{n}[\text{vec}(\hat{\xi}) - \text{vec}(\xi)] \to N(0, \Gamma)$, implying that $\sqrt{n}[\text{vec}(H^T\hat{\xi}) - \text{vec}(H^T\xi)] \to N[0, (I_{2m} \otimes H^T)\hat{\Gamma}(I_{2m} \otimes H)]$ by Slutsky's theorem. This asymptotic distribution can be used to construct the Wald test statistics for the null hypothesis $H^T\xi = 0$,

$$T(\mathcal{H}) = n\text{vec}(H^T\hat{\xi})^T[(I_{2m} \otimes H^T)\hat{\Gamma}(I_{2m} \otimes H)]^{-1}\text{vec}(H^T\hat{\xi}),$$

which is a χ^2_{2rm} asymptotically. Let $A = (I_{2m} \otimes H^T)\hat{\Gamma}(I_{2m} \otimes H)$, then the rank $(A^{1/2}A^{-1}A^{1/2}) = 2rm$, the df. In addition, $T(\mathcal{H})$ is invariant with respect to the choice of basis of $\mathcal{H}$ and can be applied to test $\mathbf{Y} \perp\!\!\!\perp X_k|\mathbf{X}_{-k}$, where X_k is the k-th predictor and $\mathbf{X}_{-k}$ indicates the remaining predictors after taking away X_k (Cook and Ni, 2005).

2. Joint dimension predictor hypotheses, $H_0 : H^T\xi = 0$ and $d = t$ vs. $H_a : H^T\xi \neq 0$ or $d > t$, test on both predictors and $d = t$. Combing information of the predictor and dimension parts, we rewrite $\xi = Q_{\mathcal{H}}\xi = Q_{\mathcal{H}}\beta v = H_0\beta_{H_0}v$, where $\beta \in \mathbb{R}^{p\times t}, v \in \mathbb{R}^{t\times 2m}$, and the coordinates $\beta_{H_0} \in \mathbb{R}^{(p-r)\times t}$ of β in terms of the basis H_0 for Span$(Q_{\mathcal{H}})$. The constrained optimal discrepancy function under a joint hypothe-

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

sis is

$$F_{t,H}(B,C) = [\text{vec}(\hat{\xi}) - \text{vec}(H_0 BC)]^T \hat{\Gamma}^{-1} [\text{vec}(\hat{\xi}) - \text{vec}(H_0 BC)].$$

The test statistic is defined as $n\hat{F}_{t,H}(B,C)$, which is asymptotically a $\chi^2_{(p-t)(2m-t)+tr}$. Because the Jacobian matrix is $\Delta_{\xi,H} = (I_{2m} \otimes H_0)(v^T \otimes I_{p-r}, I_{2m} \otimes \beta H_0) \in \mathbb{R}^{2pm \times t(p-r+2m)}$, so its df is $2pm - rank(\Delta_{\xi,H}) = 2pm - t(p-r-t+2m) = (p-t)(2m-t) + tr$.

3. Conditional Predictor Hypotheses, $H_0 : d = t$ vs $H_a : d > t$ given $H^T \xi = 0$, test on the specific dimension given that the user-specify subspace is independent of the central subspace. The test statistic is the difference of two discrepancy functions:

$$T(\mathcal{H}|d) = nF_{d,H}(B,C) - nF_d(B,C;\hat{\Gamma}^{-1}).$$

Here, $\hat{T}(\mathcal{H}|d) \sim \chi^2_{rd}$ under the null hypothesis. In fact, $T(\mathcal{H}|d)$ is asymptotically equivalent to $U^T(P_\xi - P_{\xi,H})U$, where $U \in \mathbb{R}^{2pm}$ is a standard normal random vector, and P_ξ and $P_{\xi,H}$ are the projections with respect to the usual inner product onto $\text{Span}(\Gamma^{-1/2}\Delta)$ and $\text{Span}(\Gamma^{-1/2}\Delta_{\xi,H})$, respectively. It can be shown that $\text{Span}(\Delta_{\xi,H}) \subseteq \text{Span}(\Delta)$, and thus $\text{Span}(\Gamma^{-1/2}\Delta_{\xi,H}) \subseteq \text{Span}(\Gamma^{-1/2}\Delta)$. Then $(P_\xi - P_{\xi,H})$ is a projection with $\text{rank}(\Delta) - \text{rank}(\Delta_{\xi,H}) = d(p-d+2m) - d(p-r-d+2m) = rd$.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

3. Suboptimal Estimators

Note that $pm(2pm-1)$ parameters in the limiting covariance of $\hat{\xi}$ (Theorem 1) will increase the complexity of computation. In this section, we investigate other suboptimal estimators that employ different inner product matrices, such as a diagonal block covariance and a robust version of the limiting covariance.

3.1 Degenerated Estimator (FT-DIRE)

Previously, we introduced an optimal estimator FT-IRE in an intuitive way based on the asymptotic limiting distribution. However, FT-IRE may encounter a singular covariance matrix especially when m is too large. Alternatively, diagonal block inner product matrices reduce the number of parameters to alleviate the singular problem. Or equivalently, we repeatedly construct several QDFs with their limiting covariance matrices and regard them independent. Let K be the number of QDFs. The l-th QDF has the size of $\boldsymbol{\omega}$s m_l corresponding to the limiting covariance matrix $\Gamma_l \in \mathbb{R}^{2pm_l \times 2pm_l}$, where $l = 1, ..., K$ and $\sum_{l=1}^{K} m_l = m$. The degenerated QDF is defined as the summation of the K QDFs, that is

$$F_d(B, C; \{V_l\}) = \sum_{l=1}^{K} [\mathrm{vec}(\hat{\xi}_l) - \mathrm{vec}(BC_l)]^T V_l [\mathrm{vec}(\hat{\xi}_l) - \mathrm{vec}(BC_l)], \quad (3.2)$$

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

which is equivalent to (2.1) with a new inner product matrix $V = \hat{\Gamma}_D^{-1} = \text{diag}(\{\hat{\Gamma}_l^{-1}\})$. We denote the degenerated estimator $\hat{\beta}$ via minimizing (3.2) as a Fourier Transform Degenerated Inverse Regression Estimator (FT-DIRE). Compared to FT-IRE, FT-DIRE is computationally cheaper because the inner product matrix is structurally simpler with fewer parameters to estimate. Similar to FT-IRE, it is easy to prove the consistency of FT-DIRE and the asymptotic distribution of its MDF as the following.

Theorem 3. *Assume that* $\{\mathbf{y}_k, \mathbf{x}_k\}$, $k = 1, ..., n$, *are random samples on* $(\mathbf{Y}, \mathbf{X})$ *with finite fourth moments. Let* $(\hat{\beta}, \hat{\nu}) = \arg\min_{B,C} F_d(B, C; \hat{\Gamma}_D^{-1})$. *We have*

1. $\text{Span}(\hat{\beta})$ *is a consistent estimator of* $\mathcal{S}_\xi$.

2. *As* $n \to \infty$, $\hat{\Lambda}_d = n\hat{F}_d(B, C; \hat{\Gamma}_D^{-1}) \xrightarrow{D} \sum_{k=1}^{(p-d)(2m-d)} \lambda_k C_k$. *Here,* C_k*s are independent chi-squared random variables each with one degree of freedom and* $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{(p-d)(2m-d)}$ *are eigenvalues of the covariance matrix* $Q_\Phi \Omega Q_\Phi$, *where* Φ *and* Ω *are defined in the supplementary file.*

Note that, Theorem 3 allows us to test the structural dimension d and it indicates a consistent estimator of a basis of $\mathcal{S}_\xi$.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

3.2 Special Estimator (FT-SIRE)

Section 2.1 indicates that FT is based on the generalized singular value decomposition of a kernel matrix K. It is also a sub-optimal estimator with $V = \text{diag}\{\hat{\Sigma}\}$ as in the following lemma:

Lemma 1. *Assume that* $\{\mathbf{y}_k, \mathbf{x}_k\}$, $k = 1, \cdots, n$, *are random samples on* $(\mathbf{Y}, \mathbf{X})$ *with finite fourth moments, let* $\hat{u}_1, \cdots, \hat{u}_p$ *be the eigenvectors of* K *corresponding to eigenvalues* $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$. *Let* $(\hat{\beta}, \hat{\nu}) = \arg\min_{B,C} F_d(B, C; \text{diag}\{\hat{\Sigma}\})$. *Then* $\text{Span}(\hat{\beta})$ *is equal to* $\text{Span}(\hat{u}_1, \cdots, \hat{u}_d)$.

We denote $\hat{\beta}$ as Fourier Transform Special Inverse Regression Estimator (FT-SIRE). The special QDF is written as:

$$\begin{aligned} F_d(B, C; \text{diag}\{\hat{\Sigma}\}) &= [\text{vec}(\hat{\xi}) - \text{vec}(BC)]^T \text{diag}\{\hat{\Sigma}\}[\text{vec}(\hat{\xi}) - \text{vec}(BC)] \\ &= \textstyle\sum_{l=1}^{2m} (\hat{\xi}_l - BC_l)^T \hat{\Sigma} (\hat{\xi}_l - BC_l), \end{aligned} \tag{3.3}$$

where each column of ξ is considered as independent of each other. Then we can state the consistency of estimate and QDF's asymptotic distribution:

Theorem 4. *Assume that* $\{\mathbf{y}_k, \mathbf{x}_k\}$, $k = 1, \cdots, n$, *are random samples on* $(\mathbf{Y}, \mathbf{X})$ *with finite fourth moments. Let* $(\hat{\beta}, \hat{\nu}) = \arg\min_{B,C} F_d(B, C; \text{diag}\{\hat{\Sigma}\})$. *We have*

1. $\text{Span}(\hat{\beta})$ *is a consistent estimator of* $\mathcal{S}_\xi$.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

2. *As* $n \to \infty$, $\hat{\Lambda}_d = n\hat{F}_d(B, C; \text{diag}\{\hat{\Sigma}\}) \xrightarrow{D} \sum_{k=1}^{(p-d)(2m-d)} \lambda_k C_k$.

Both Theorems 3 and 4 indicate that the corresponding test statistic $\hat{\Lambda}_d$ follows a weighted χ^2 instead of a χ^2 as in Theorem 2. However, $\hat{\Lambda}_d$ can have the same asymptotic χ^2 distribution under the marginal covariance condition (Cook, 1998b), which simplifies $Q_\Phi \Omega Q_\Phi$.

3.3 Robust Estimators (FT-RIRE)

To achieve a consistent estimate of $\Gamma^{-1} = \text{Cov}\{\text{vec}[\Sigma^{-1/2}\mathbf{Z}\boldsymbol{\epsilon}^T]\}^{-1}$, we needs to plug in consistent sample estimate of Γ, which involves fourth moments of the predictors. We consider a robust estimator and first let the covariance matrix Σ be known. Let $\widetilde{\boldsymbol{\xi}}_j = \Sigma^{-1}(\frac{1}{n}\sum_{k=1}^n e^{i\boldsymbol{\omega}_j^T \mathbf{y}_k}\mathbf{x}_k - \frac{1}{n}\sum_{k=1}^n e^{i\boldsymbol{\omega}_j^T \mathbf{y}_k}\bar{\mathbf{x}})$, and $\widetilde{\epsilon} = e^{i\boldsymbol{\omega}^T\mathbf{Y}} - \text{E}e^{i\boldsymbol{\omega}^T\mathbf{Y}}$. Define $\widetilde{\xi} = \left(\widetilde{\boldsymbol{\xi}}_j^R, \widetilde{\boldsymbol{\xi}}_j^I\right)_{j=1}^m \in \mathbb{R}^{p\times 2m}$.

Theorem 5. *Assume that* $\{\mathbf{y}_k, \mathbf{x}_k\}$, $k = 1, ..., n$, *are random samples on* $(\mathbf{Y}, \mathbf{X})$ *with finite second moments. Let* $\widetilde{\boldsymbol{\epsilon}} = (\widetilde{\epsilon}_1^R, \widetilde{\epsilon}_1^I, ..., \widetilde{\epsilon}_m^R, \widetilde{\epsilon}_m^I)^T$, *where* $\widetilde{\epsilon}_j^R, \widetilde{\epsilon}_j^I$ *are the real and imaginary parts of* $\widetilde{\epsilon}_j$ *for* $j = 1, ..., m$. *Then*

$$\sqrt{n}[\text{vec}(\widetilde{\xi}) - \text{vec}(\beta\nu)] \xrightarrow{D} N(0, \widetilde{\Gamma}),$$

where $\widetilde{\Gamma} = (I \otimes \Sigma^{-1/2})\text{Cov}[\text{vec}(\mathbf{Z}\widetilde{\boldsymbol{\epsilon}}^T)](I \otimes \Sigma^{-1/2})$.

The proof of Theorem 5 follows from Theorem 1, so we omit it. The resulting limiting covariance matrix is $\widetilde{\Gamma} = (I \otimes \Sigma^{-1/2})\text{Cov}[\text{vec}(\mathbf{Z}\widetilde{\boldsymbol{\epsilon}}^T)](I \otimes$

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

$\Sigma^{-1/2}$), which only requires the second moments of the preditor. Also, computing its inverse $\widetilde{\Gamma}^{-1} = (I \otimes \Sigma^{1/2})\text{Cov}[\text{vec}(\mathbf{Z}\widetilde{\boldsymbol{\epsilon}}^T)]^{-1}(I \otimes \Sigma^{1/2})$ is structurally simplier and computationally cheaper than $\Gamma^{-1} = \text{Cov}\{\text{vec}[\Sigma^{-1/2}\mathbf{Z}\boldsymbol{\epsilon}^T]\}^{-1}$. Because Γ^{-1} involves estimation of the predictor' fourth moments, while $\widetilde{\Gamma}^{-1}$ only needs its second moments, FT-RIRE is more theoretically robust.

Theorem 5 shows the limiting covariance matrix $\widetilde{\Gamma}$ of $\widetilde{\xi}$, so it is natural to define the robust QDF as

$$F_d(B, C; \widetilde{G}^{-1}) = [\text{vec}(\hat{\xi}) - \text{vec}(BC)]^T \widetilde{G}^{-1} [\text{vec}(\hat{\xi}) - \text{vec}(BC)],$$

where $\widetilde{G} = (I \otimes \hat{\Sigma}^{-1/2})\widehat{\text{Cov}}[\text{vec}(\mathbf{Z}\widetilde{\boldsymbol{\epsilon}}^T)](I \otimes \hat{\Sigma}^{-1/2})$. The estimator by minimizing the robust QDF is called as the Fourier Transform Robust Inverse Regression Estimator (FT-RIRE), which has the following properties.

Theorem 6. *Assume that* $\{\mathbf{y}_k, \mathbf{x}_k\}$,$k = 1, ..., n$, *are random samples on* $(\mathbf{Y}, \mathbf{X})$ *with finite second moments and let* $(\hat{\beta}, \hat{\nu}) = \arg\min_{B,C} F_d(B, C; \widetilde{G}^{-1})$. *Then*

1. $n\hat{F}_d(B, C; \widetilde{G}^{-1})$ *has an asymptotic chi-squared distribution with degrees of freedom* $(p - d)(2m - d)$.

2. $\text{Span}(\hat{\beta})$ *is a consistent estimator of* $\mathcal{S}_\xi$.

We can also define a diagonal block inner product matrix that is $\widetilde{G}_D^{-1} = \text{diag}\{\widetilde{G}_1^{-1}, ..., \widetilde{G}_K^{-1}\}$ following the notations in Section 3.1. The degenerated

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

robust estimator minimizing $F_d(B, C; \widetilde{G}_D^{-1})$ is named as Fourier Transform Degenerated Robust Inverse Regression Estimator (FT-DRIRE).

Theorem 7. *Assume that* $\{\mathbf{y}_k, \mathbf{x}_k\}$,$k = 1, ..., n$, *are random samples on* $(\mathbf{Y}, \mathbf{X})$ *with finite second moments. Let* $(\hat{\beta}, \hat{\nu}) = \arg\min_{B,C} F_d(B, C; \widetilde{G}_D^{-1})$. *We have*

1. $\text{Span}(\hat{\beta})$ *is a consistent estimator of* $\mathcal{S}_\xi$.

2. *As* $n \to \infty$, $\hat{\Lambda}_d = n\hat{F}_d(B, C; \widetilde{G}_D^{-1}) \xrightarrow{D} \sum_{k=1}^{(p-d)(2m-d)} \lambda_k C_k$.

We then have similar discussion for the robust estimators as Sections 3.1 and 3.2. The proofs of 6 and 7 follow straightforwardly from Theorem 3. Hence, we omit them.

In summary, FT-IRE is optimal because it is asymptotically efficient without any constraints or strong assumptions. On the other hand, the other four estimators: FT-DIRE, FT-SIRE, FT-RIRE, and FT-DRIRE, are suboptimal because we employ either a diagonal block covariance matrix assuming some independent structure or a robust version of the limiting covariance matrix. Those estimators are minimizers over corresponding constrained spaces. We will present more simulations in the next section.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

4. Numerical Study

In this section, we provide folllowing simulations to evaluate the performance of the minimum discrepancy with Fourier transform approaches. The following criterion is used to compare the accuracy between B and its estimate $\hat{B}$, where both are $p \times d$ orthogonal matrices. Trace correlation $r_2 = \sqrt{\sum_{i=1}^{d} \rho_i^2/d}$ (Ye and Weiss, 2003), where ρ_i are the eigenvalues of matrix $\hat{B}^T B B^T \hat{B}$ for $i = 1, ..., d$. Note that $r_2 \in [0, 1]$, and the bigger r_2 indicates better estimate. The Frobenius norm: $||\hat{B}\hat{B}^T - BB^T||_F$ (Li et al., 2005). The smaller Frobenius norm indicates better estimate.

4.1 Simulations

Model 1. *This model shows the effect of the different sizes of* $\boldsymbol{\omega}$ *on the performance of FT-IRE, FT-DIRE, FT-SIRE, FT-RIRE, and FT-DRIRE. We examine four examples, where examples 1.1-1.3 are the modifications of Lin et al. (2019), while 1.4 is a multivariate response example, similar to example 3 of Zhu et al. (2010). Let* $p = 10, n = 100$, $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_p)$. *For 1.1-1.3, column* β_i, *for* $i = 1, 2$, *has coefficient 1 at* 3 *random positions, otherwise, 0; and that* $\beta_i \neq \beta_j$, *for* $i \neq j$. *The error term* $\epsilon \sim N(0, 1)$. *For 1.4,* β_1 *and* β_2 *have fixed coefficients.*

1.1. $Y = (\beta_1^T \mathbf{X})^{3/2} + \epsilon$, $d = 1$.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

1.2. $Y = \exp(\beta_1^T \mathbf{X} + 0.5\epsilon),\ d = 1.$

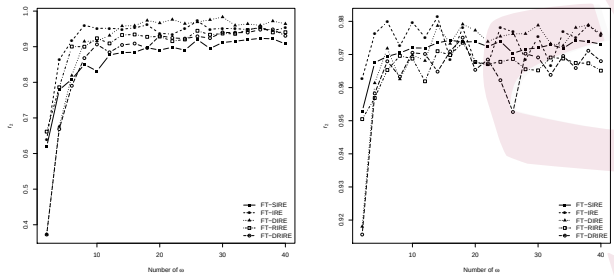
1.3. $Y = |\frac{\beta_2^T \mathbf{X}}{4} + 2|^3 sign(\beta_1^T \mathbf{X}) + \epsilon,\ d = 2.$

1.4. $Y_1 = 1 + \beta_1^T \mathbf{X} + \sin(\beta_2^T \mathbf{X}) + \epsilon_1,\ Y_2 = \frac{\beta_2^T \mathbf{X}}{0.5+(\beta_1^T \mathbf{X}+1)^2} + \epsilon_2,\ Y_3 = |\beta_1^T \mathbf{X}|\epsilon_3,$ $Y_4 = \epsilon_4, Y_5 = \epsilon_5,$ *with* $d = 2,\ \beta_1 = (1, 0, ..., 0)^T,$ *and* $\beta_2 = (0, 1, 1, 0, ..., 0)^T.$ *Error terms are* $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, ..., \epsilon_q)^T \sim N_q(\mathbf{0}, \boldsymbol{\Sigma}),$ *where* $q = 5$ *and* $\boldsymbol{\Sigma} = \begin{pmatrix} A & 0 \\ 0 & D \end{pmatrix}$ *with* $A = \begin{pmatrix} 1 & -1/2 \\ -1/2 & 1/2 \end{pmatrix}$ *and* $D = \text{diag}(1/2, 1/3, ..., 1/q).$

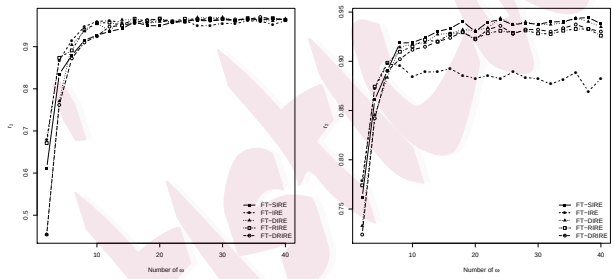
Figure 1 plots the mean values of r_2 over 100 simulated data vs. the different sizes of $\boldsymbol{\omega}$: $\{2, 4, ..., 40\}$ using the QL decomposition of soft-thresholding covariance Γ. In these four panels, r_2 values rapidly increase when the size of $\boldsymbol{\omega}$ reaches 4, which indicates that $m = 4$ has included important information. All the estimates are accurate and stable as the size of $\boldsymbol{\omega}$ sufficiently large (Here, say, larger than 2 when d is 1 or 2). For example 1.1-1.3, FT-IRE and FT-DIRE perform as good as their corresponding robust versions. While FT-SIRE is the best in accuracy for example 1.4 and it is also more stable than others for example 1.2. FT-SIRE is promising because it has the simplest structure of the inner product matrix $\Lambda = \Gamma^{-1}$, and it is also computationally faster than other methods.

Model 2. *This model is used to compare our approaches with FIRE, DIRE (Cook and Zhang, 2014) and IRE (Cook and Ni, 2005), and also compare*

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)



(a) Example 1.1 (b) Example 1.2



(c) Example 1.3 (d) Example 1.4

Figure 1: Using QL decomposition of soft-thresholding covariance: Mean values of r_2 over 100 simulated data vs. the different sizes of $\boldsymbol{\omega}$: $\{2, 4, ..., 40\}$ in Model 1.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

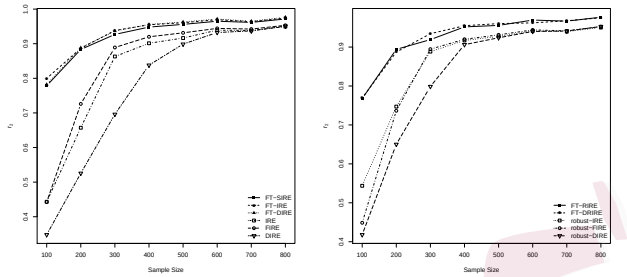
corresponding robust versions FT-RIRE, FT-DRIRE with robust FIRE, robust DIRE (Cook and Zhang, 2014), and robust IRE (Ni and Cook, 2007).

$$Y = |\mathbf{X}^T\beta| + 0.2\epsilon,$$

$p = 10, d = 1$, *and* $\beta = (1, 1, 1, 1, 0, \cdots, 0)^T \in \mathbb{R}^p$. *Predictors* $\mathbf{X} \sim N(\boldsymbol{\mu}_s, \Sigma)$, *where* $\boldsymbol{\mu}_s \in \mathbb{R}^p$ *is a p-dimensional vector with a value* 2 *at one random place of* $j = 1, \cdots, p$ *and* 0 *otherwise, and* Σ *has a first-order autoregressive structure with* $(j_1, j_2)th$ *entry* $0.5^{|j_1 - j_2|}$. *Sample sizes range from 100 to 800 with increment 100. We use size m of* $\boldsymbol{\omega}$ *4. Other values* $m = 6, 10, 16$ *are reported in the supplementary file (Figure S2.4). There is no difference in what value m to use because* $m = 4$ *produces stable results for this model. Settings for existing methods are below unless otherwise specified: the fused slices for FIRE and DIRE are* $H = \{3, 4, ..., 15\}$, *and the slice number for IRE is* $h = 5$.

The left panel of Figure 2 compares our approaches, FT-IRE , FT-DIRE , and FT-SIRE with existing methods, FIRE, DIRE, and IRE. Our approaches outperform FIRE, DIRE, and IRE in terms of the higher r_2 values. As expected, larger the sample size produces better estimates of all these methods. Even for sample size is as small as $n = 100$, and our approaches have larger r_2 than the other three slicing methods.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)



(a) Number of ω: 4 (b) Number of ω: 4

Figure 2: Mean values of r_2 over 100 simulated data vs. various sample sizes from 100 to 800 at increment 100 in Model 2.

The right panel of Figure 2 shows the comparisons between our robust estimators, FT-RIRE, FT-DRIRE with robust FIRE, robust DIRE (Cook and Zhang, 2014), and robust IRE (Ni and Cook, 2007). Overall, our methods are better than others in terms of estimation accuracy, especially in smaller sample sizes.

Model 3. *This model evaluates the asymptotic performance of our proposed methods on dimension detection compared with FT (Weng and Yin, 2018). We examine four examples. Examples 3.1-3.2 are the modifications of the examples in Lin et al. (2019), and predictor vector* $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_p)$ *with* $p = 10$*, and dimension* $d = 2$*. The column* β_i*, for* $i = 1, 2$*, has coefficient 1 at 3 random positions, otherwise, 0; and that* $\beta_i \neq \beta_j$*, for* $i \neq j$*. Error term* $\epsilon \sim N(0, 1)$*. While for example 3.3,* β_1 *and* β_2 *have fixed coefficients.*

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

This example comes from Weng and Yin (2018) involves certain degree of collinearity. Predictors $X_1, X_3, X_5, \cdots X_p \overset{iid}{\sim} N(0,1)$, *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* i^{th} *entry is 1 and other entries are 0. Then* $(\beta_1, \beta_2) = (e_1, e_2)$. *The last example 3.4 is the same as example 1.4.*

3.1. $Y = (\beta_1^T \mathbf{X}) \exp(\beta_2^T \mathbf{X}) + \epsilon$.

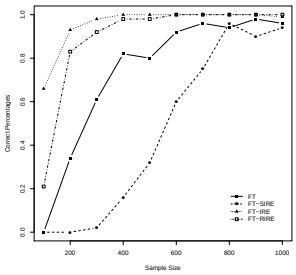
3.2. $Y = (\beta_1^T \mathbf{X}) \exp(\beta_2^T \mathbf{X} + \epsilon)$.

3.3. $Y = X_1 + 0.5X_2^2$.

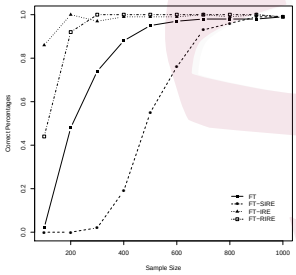
3.4. This is the same as example 1.4.

Figure 3 shows the percentages of correctly detecting dimensions ($d = 2$) over 100 simulated data for different sample size: $\{100, ..., 1000\}$ using QL decomposition of soft-thresholding covariance with $m = 4$. Overall, FT-IRE and FT-RIRE have higher correct rates than the other methods. However, correct rates do not increase for degenerated estimators FT-DIRE and FT-DRIRE as the sample size increases (see Figure S2.5 in the supplementary file). We don't recommend using the degenerated estimators for structural dimension and predictor hypothesis tests due to their strong assumption of the discrepancy functions.

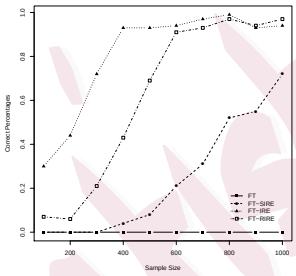
Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)



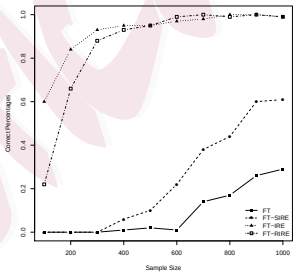
(a) Example 3.1



(b) Example 3.2



(c) Example 3.3



(d) Example 3.4

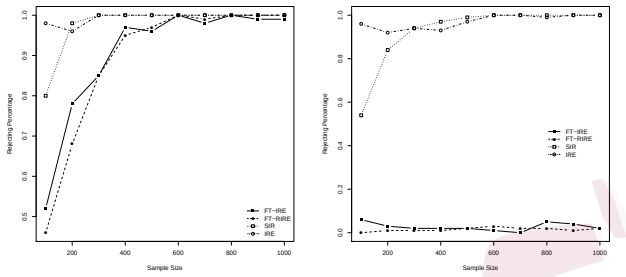
Figure 3: Percentages of correctly detecting dimensions ($d = 2$) over 100 simulated data vs. sample sizes n: $\{100, ..., 1000\}$ in Model 3.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

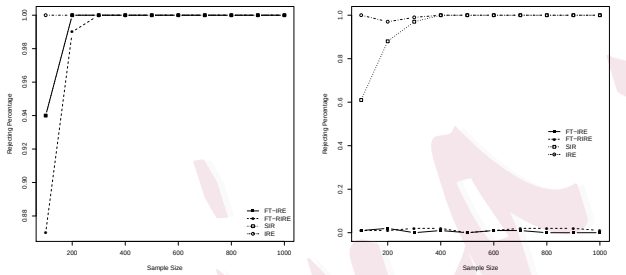
Almost all inverse approaches depending on $\mathrm{E}(\mathbf{X}|\mathbf{Y})$ can not detect the symmetric link function. Fortunately, both Example 3.3 and 3.4 contain two linear terms, which can be obtained using the inverse approaches. Especially, Example 3.3 can be written as $Y = X_1 + 0.5X_4 + 0.5X_1^2 + 0.5(X_1 - 1)Z$. Other symmetric components produce more noises raising the challenges in testing. The test statistic's distribution for FT and FT-SIRE has a weighted chi-squared distribution. Li (1998) noted that the weights are nonzero eigenvalues of some nonnegative definite symmetric matrix, and the variances of weights would most likely be sizable. The estimation of weights has a significant impact on the performance of dimension detection. It turns out that the test statistics' distributions for FT-SIRE and FT are not as accurate as FT-IRE and FT-RIRE, especially when the signal is not strong enough. Figure 3 (c, d) indicates that our proposed methods can unearth two directions from noises, but FT fails in most cases.

Model 4. *This model investigates the effect of the sample size on the performance of marginal predictor hypothesis tests. We only consider FT-IRE and FT-RIRE compared with SIR and IRE for testing user-specified subspaces. Similar to structural dimension testings, FT-SIRE, FT-DIRE, and FT-DRIRE are not suggested because of their assumptions on the quadratic functions. Predictor* $p \times 1$ *vector* $\mathbf{X} \sim N(0, I)$ *with* $p = 10$ *and* $n = 100$.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)



(a) Example 4.1 with $\mathcal{H}_1$ (b) Example 4.1 with $\mathcal{H}_2$



(c) Example 4.2 with $\mathcal{H}_1$ (d) Example 4.2 with $\mathcal{H}_2$

Figure 4: Percentages of rejecting using Marginal predictors hypothesis tests with $m = 4$ in Model 4.

There are two examples 4.1-4.2, column β_i*, for* $i = 1, 2$*, has coefficient 1 at* 3 *random positions, otherwise, 0; and that* $\beta_i \neq \beta_j$*, for* $i \neq j$*. Error term* $\epsilon \sim N(0, 1)$*. Let* $d = 2$ *and use* $m = 4$*:*

4.1. $Y = \beta_1^T \mathbf{X}(\beta_2^T \mathbf{X} + 1) + \epsilon,$

4.2. $Y = \frac{\beta_1^T \mathbf{X}}{0.5+(\beta_2^T \mathbf{X}+1.5)^2} + \epsilon.$

Two user-specified subspaces are $\mathcal{H}_1 \in \mathbb{R}^{p\times 1}$ with 1s in the rows of the

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

same nonzero rows of β and 0s otherwise, and $\mathcal{H}_2 \in \mathbb{R}^{p\times 1}$ with 1s in the rows of the same zero rows of β and 0s otherwise. The percentages of rejecting the null hypothesis H_0: $\mathcal{H}^T\beta = 0$ ($\mathcal{H}$ represents $\mathcal{H}_1$ or $\mathcal{H}_2$) are presented among 100 simulated data, given the significant level 0.05. The rejection rate for $\mathcal{H}_1$ indicates the power, while the rejection rate for $\mathcal{H}_2$ means Type I Error.

The left two panels of Figure 4 show that the power of our marginal predictor tests is close to one when the sample size varies from 500 to 1000. The right two panels of Figure 4 illustrate that the empirical Type I Error rates of FT-IRE or FT-RIRE in the marginal predictor hypothesis tests are under reasonable control, below 0.05. While SIR and IRE have huge Type I Error rates, close to 1. As long as the sample size is larger than 500, our tests are superior with smaller Type I Errors, supporting asymptotic results.

4.2 Data analysis

We use the Prostate data (Stamey et al., 1989) to compare FT-IRE and FT-DIRE with SIR, IRE, FIRE, and DIRE. The data studies the level of a prostate-specific antigen, associated with eight clinical measures in 97 male patients taking a radical prostatectomy. It is availabe at rafalab.github.io.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

The eight clinical measurements are: logarithm of cancer volume (lcavol) and of prostate weight (lweight), age (age), logarithm of benign prostatic hyperplasia amount (lbph), seminal vesicle invasion (svi), logarithm of capsular penetration (lcp), Gleason score (gleason), and percentage Gleason scores 4 or 5 (pgg45). The outcome is the logrithm of prostate specific antigen (lpsa).

Testing the structural dimension using SIR and FT-IRE result in $d = 1$, while IRE results in $d = 2$. We use both estimated d to illustrate our points. Furthermore, we fit six regression models (with the respective estimated d): Nadaraya-Watson kernel regression, local linear regression, polynomial regression with degree 2, polynomial regression with degree 3, generalized additive model, and linear regression. For each model and method, we calculate the mean squared errors (MSE) using the 5-folds cross-validation, shown in Table 1. We can conclude that our FT-IRE is the best among all these comparisons.

To further assess the performance of these six methods, we use the IRE estimate with $d = 2$, then fit a polynomial model of order 2 without the intersection term because it is insignificant (p-value is less than 0.05), resulting in a $\hat{\sigma}$ and $\hat{Y} = \alpha_1 X_1^* + \alpha_2 X_2^* + \alpha_{11} X_1^{*2} + \alpha_{22} X_2^{*2}$. We generate another 100 data sets using $Y = \hat{Y} + \epsilon$, with different sample sizes 50, 100,

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Table 1: The MSEs with six regression models for each method.

	$d=2$			$d=1$		
Regression Models	IRE	FIRE	DIRE	SIR	FT-IRE	FT-DIRE
Nadaraya-Watson	0.9293	0.8088	0.7979	0.6971	0.6941	0.6944
Local Linear	0.8585	0.8264	0.7822	0.7341	0.7348	0.7348
Polynomial (dg=2)	0.8697	0.8778	0.9158	0.7205	0.7170	0.7200
Polynomial (dg=3)	0.9415	0.8265	0.9420	0.7518	0.7488	0.7495
GAM	0.9198	0.7804	0.9413	0.7549	0.7360	0.7412
Linear Regression	0.8699	0.7790	0.8402	0.7070	0.7037	0.7046

200, and 400, where $\epsilon \sim N(0, \hat{\sigma})$. We evaluate the estimation accurate. Table 2 indicates that the Fourier transform approaches have the higher accuracy in estimation (smaller in MSE and Frobenius Norm, and higher in r^2) compared to IRE, FIRE and DIRE, and slightly better than SIR.

5. Discussion

In this paper, we developed an optimal minimum discrepancy with Fourier transform approach in sufficient dimension reduction, which is especially useful in the multivariate scenario without a slicing scheme of the response. There are four sub-optimal estimators are discussed for computational efficiency and robustness.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Table 2: Comparing six methods using IRE estimation with $d = 2$.

n	Criteria	IRE	FIRE	DIRE	SIR	FT-IRE	FT-DIRE
50	MSE	1.0716	0.9755	1.0194	0.8311	0.8193	0.8208
	r_2	0.5870	0.5872	0.5460	0.6404	0.6663	0.6604
	Norm	1.5997	1.5895	1.6487	1.5112	1.4657	1.4764
100	MSE	1.0353	0.8392	0.9196	0.8121	0.8068	0.8069
	r_2	0.6146	0.6680	0.6121	0.7041	0.6929	0.6900
	Norm	1.5617	1.4670	1.5501	1.4045	1.4269	1.4319
200	MSE	1.0331	0.8313	0.8881	0.8245	0.8228	0.8228
	r_2	0.6017	0.6933	0.7088	0.7114	0.7147	0.7137
	Norm	1.5818	1.4283	1.3902	1.3932	1.3878	1.3902
400	MSE	1.0142	0.8349	0.8442	0.8295	0.831	0.8292
	r_2	0.6192	0.7131	0.7074	0.7393	0.7414	0.7409
	Norm	1.5468	1.3924	1.4047	1.3353	1.3312	1.3319

Among all five proposed methods, we recommend firstly considering FT-IRE in the real application. From a theoretical perspective, FT-IRE is asymptotic efficient, and so optimal. Empirically, FT-IRE not only provides competitive estimation as the other four methods, but it also has a better performance in testing structural dimension and significant predictor. In comparison, FT-DIRE and FT-SIRE are superior in scenarios with a limited computational environment. FT-RIRE and FT-DRIRE are better for a

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

small sample size.

Besides, we develop marginal, conditional predictors tests, and joint tests for FT-IRE. We demonstrate the effectiveness and usefulness of our proposed approaches. Furthermore, the minimum discrepancy with Fourier transform approach can be extended to partial sufficient dimension reduction (Chiaromonte et al., 2002; Li et al., 2003) when dealing with categorical and numeric variables. We believe that it is an exciting research direction.

Supplementary Materials

The Supplementary Material contains the proofs of theorems and detailed tables and figures for the simulation studies.

Acknowledgements

This work is supported in part by an NSF grant CIF 1813330. The authors would like to thank the Editor, an Associate Editor and two referees for their valuable comments and suggestions, which lead to a greatly improved paper.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

References

Bickel, P. J. and E. Levina (2008). Covariance regularization by thresholding. *The Annals of Statistics 36*(6), 2577–2604.

Chiaromonte, F., R. D. Cook, and B. Li (2002). Sufficient dimension reduction in regressions with categorical predictors. *The Annals of Statistics 30*(2), 475–497.

Cook, R. D. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association 91*(435), 983–992.

Cook, R. D. (1998a). Principal hessian directions revisited. *Journal of the American Statistical Association 93*(441), 84–94.

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

Cook, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *The Annals of Statistics 32*(3), 1062–1092.

Cook, R. D. and L. Ni (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Association 100*(470), 410–428.

Cook, R. D. and C. M. Setodji (2003). A model-free test for reduced rank in multivariate regression. *Journal of the American Statistical Association 98*(462), 340–351.

Cook, R. D. and S. Weisberg (1991). Comment on "Sliced inverse regression for dimension reduction" by K.-C. Li. *Journal of the American Statistical Association 86*(414), 328–332.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

REFERENCES

Cook, R. D. and X. Zhang (2014). Fused estimators of the central subspace in sufficient dimension reduction. *Journal of the American Statistical Association 109*(506), 815–827.

Li, B. (2018). *Sufficient dimension reduction: Methods and applications with R*. CRC Press.

Li, B., R. D. Cook, and F. Chiaromonte (2003). Dimension reduction for the conditional mean in regressions with categorical predictors. *The Annals of Statistics 31*(5), 1636–1668.

Li, B. and S. Wang (2007a). On directional regression for dimension reduction. *Journal of the American Statistical Association 102*(479), 997–1008.

Li, B. and S. Wang (2007b). On directional regression for dimension reduction. *Journal of the American Statistical Association 102*(479), 997–1008.

Li, B., S. Wen, and L. Zhu (2008). On a projective resampling method for dimension reduction with multivariate responses. *Journal of the American Statistical Association 103*(483), 1177–1186.

Li, B., H. Zha, and F. Chiaromonte (2005). Contour regression: A general approach to dimension reduction. *The Annals of Statistics 33*(4), 1580–1616.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association 86*(414), 316–327.

Li, K.-C. (1998). Principal hessian directions revisited: Comment. *Journal of the American Statistical Association 93*(441), 94–97.

Lin, Q., Z. Zhao, and J. Liu (2019). Sparse sliced inverse regression via lasso. *Journal of the*

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

American Statistical Association 114(528), 1726–1739.

Ni, L. and R. D. Cook (2007). A robust inverse regression estimator. *Statistics and Probability Letters 77*(3), 343–349.

Qian, W., S. Ding, and R. D. Cook (2019). Sparse minimum discrepancy approach to sufficient dimension reduction with simultaneous variable selection in ultrahigh dimension. *Journal of the American Statistical Association 114*(527), 1277–1290.

Rothman, A. J., E. Levina, and J. Zhu (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association 104*(485), 177–186.

Saracco, J. (2005). Asymptotic for pooled marginal slicing estimator based on sir_α approach. *Journal of Multivariate Analysis 96*(1), 117–135.

Shapiro, A. (1986). Asymptotic theory of overparameterized structural models. *Journal of the American Statistical Association 81*(393), 142–149.

Stamey, T. A., J. N. Kabalin, J. E. McNeal, I. M. Johnstone, F. Freiha, E. A. Redwine, and N. Yang (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. ii. radical prostatectomy treated patients. *The Journal of urology 141*(5), 1076–1083.

Wang, H. and Y. Xia (2008). Sliced regression for dimension reduction. *Journal of the American Statistical Association 103*(482), 811–821.

Weng, J. and X. Yin (2018). Fourier transform approach for inverse dimension reduction method.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Journal of Nonparametric Statistics 30(4), 1049–1071.

Ye, Z. and R. E. Weiss (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association 98*(464), 968–979.

Yin, X. and E. Bura (2006). Moment-based dimension reduction for multivariate response regression. *Journal of Statistical Planning and Inference 136*(10), 3675–3688.

Yin, X. and H. Hilafu (2015). Sequential sufficient dimension reduction for large p, small n problems. *Journal of the Royal Statistical Society, Series B 77*, 879–892.

Yin, X., B. Li, and R. D. Cook (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis 99*(8), 1733–1757.

Zhu, L., L. Zhu, and S. Wen (2010). On dimension reduction in regressions with multivariate responses. *Statistica Sinica 20*(1), 1291–1307.

Zhu, Y. and P. Zeng (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *Journal of the American Statistical Association 101*(476), 1638–1651.

Mathematical Sciences, Bentley University, Waltham, MA 02452, U.S.A.

E-mail: jweng@bentley.edu

Department of Statistics, University of Kentucky, Lexington, KY 40536, U.S.A.

E-mail: yinxiangrong@uky.edu