Fourier Methods for Estimating The Central Subspace and The Central Mean Subspace in Regression

Yu Zhu and Peng Zeng

Purdue University and University of Auburn

Abstract

In high dimensional regression, it is important to estimate the central and central mean subspaces, to which the projections of the predictors preserve sufficient information about the response and the mean response, respectively. Using the Fourier transform, we have derived the candidate matrices whose column spaces recover the central and central mean subspaces exhaustively. Under the normality assumption of the predictors, explicit estimates of the central and central mean subspaces are derived. Bootstrap procedures are used for determining dimensionality and choosing tuning parameters. Simulation results and an application to a real data are reported. Our methods demonstrate competitive performance compared to SIR, SAVE and other existing methods. The approach proposed in the paper provides a novel view on sufficient dimension reduction and may lead to more powerful tools in the future.

KEY WORDS: Central subspace; Central mean subspace; SIR; SAVE; Candidate matrix; Fourier transform; Bootstrap

1 Introduction

Suppose Y is a univariate response and X is a p-dimensional vector of continuous predictors. Let $F_{Y|X}$ denote the conditional distribution of Y given X and $E[Y|X]$ the mean response at X. In full generality, the regression of Y on X is to infer about the conditional distribution $F_{Y|X}$ often with the mean response $E[Y|X]$ of primary interest. When $F_{Y|X}$ or $E[Y|X]$ does not admit a proper parametric form, nonparametric methods such as local polynomial smoothing are usually employed for regression. Due to the curse of dimensionality, however, these methods become impractical when the dimension of X is high. In order to mitigate the curse of dimensionality, various dimension reduction techniques have been proposed in the literature. One popular approach is to project X onto a lower dimensional subspace where the regression of Y on X can be performed. In this paper, we focus on *sufficient dimension reduction* (SDR), which further requires that the projection of X onto the lower dimensional subspace does not result in any loss of information about $F_{Y|X}$ or $E[Y|X]$.

The theory of sufficient dimension reduction was originated from the seminal works by Li (1991) and Cook and Weisberg (1991). During the past decade, much progress has been achieved in SDR; see Cook (1998) for a comprehensive account. Let $\mathcal{S}$ denote a subspace of $\mathbb{R}^p$ and $P_{\mathcal{S}}$ be the orthogonal

projection operator onto $\mathcal{S}$ in the usual inner product. $\mathcal{S}$ is called a dimension reduction subspace if Y and X are independent conditioned on $P_{\mathcal{S}}X$, that is

$$Y \perp\!\!\!\perp X \mid P_{\mathcal{S}}X, \tag{1}$$

where $\perp\!\!\!\perp$ means "independent with". Note that dimension reduction subspace may not be unique. When the intersection of all dimension reduction subspaces is also a dimension reduction subspace, it is defined to be the *central subspace*, denoted by $\mathcal{S}_{Y|X}$ (Cook 1996, 1998). The dimension of $S_{Y|X}$ is called the *structural dimension* of the regression of Y on X, which is denoted by $\dim(\mathcal{S}_{Y|X})$. $\mathcal{S}_{Y|X}$ can be regarded as a metaparameter that is the target of sufficient dimension reduction for $F_{Y|X}$. Under mild conditions, Cook (1996, 1998) showed that central subspace exists and is unique. Throughout this paper, we assume the existence of the central subspace.

When only the mean response $E[Y|X]$ is of interest, sufficient dimension reduction can be defined for $E[Y|X]$ in a similar fashion as for $F_{Y|X}$. A subspace $\mathcal{S}$ is called a mean dimension reduction subspace if

$$Y \perp\!\!\!\perp E[Y|X] \mid P_{\mathcal{S}}X. \tag{2}$$

If the intersection of all mean dimension reduction subspaces is also a mean dimension reduction subspace, it is defined to be the *central mean subspace* denoted by $\mathcal{S}_{E[Y|X]}$ (Cook and Li 2002). Similar to the central subspace, the central mean subspace exists under mild conditions; so its existence is assumed throughout this paper. $\mathcal{S}_{E[Y|X]}$ is the target of sufficient dimension reduction for the mean response $E[Y|X]$ and it is always a subspace of the central subspace $\mathcal{S}_{Y|X}$ (Cook and Li 2002). Lately, Yin and Cook (2002) extended the central mean subspace to the central kth moment subspace that is sufficient for the first k moments of the conditional distribution $F_{Y|X}$.

A variety of dimension reduction methods have been proposed in the literature, some of which can be used to estimate central subspace or central mean subspace. For central subspace, they include sliced inverse regression (SIR; Li 1991) and sliced average variance estimation (SAVE; Cook and Weisberg 1991); for central mean subspace, they include principal Hessian direction (pHd; Li 1992), iterative Hessian transformation (IHT; Cook and Li 2002), structure adaptive method (SAM; Hristache et al. 2001) and minimum average variance estimation (MAVE; Xia et al. 2002). SAM and MAVE are fundamentally different from the other methods mentioned above in that both of them involve nonparametric estimation of the link function $E[Y|X = x]$, which could be impractical when the dimension of X is high. All the other methods mentioned above avoid high dimensional nonparametric estimation and target either $S_{Y|X}$ or $S_{E[Y|X]}$ directly. They usually follow a common procedure

consisting of two steps: The first step is to define a $p \times p$ nonnegative definite matrix M called candidate matrix (Ye and Weiss 2003), whose columns span a subspace of $S_{E[Y|X]}$ or $S_{Y|X}$, and then propose a consistent estimate $\hat{M}$ of the candidate matrix from a sample $\{(x_i, y_i)\}_{1 \le i \le n}$ of (X, Y); The second step is to obtain the spectral decomposition of $\hat{M}$ and use the space spanned by the eigenvectors of $\hat{M}$ corresponding to the largest q eigenvalues as the estimate of $S_{E[Y|X]}$ or $S_{Y|X}$, where q is the dimension of $S_{E[Y|X]}$ or $S_{Y|X}$. Recently, a more efficient method called minimum discrepancy method was proposed by Cook and Ni (2005) for estimating $S_{E[Y|X]}$ or $S_{Y|X}$ from a given candidate matrix. For these methods to work, some distributional assumptions need to be imposed on X. For convenience, we assume that the mean of X is the origin of $\mathbb{R}^p$ and the covariance matrix of X is the standard $p \times p$ identity matrix I_p. Then, SIR and IHT require X to satisfy the linearity condition, which is $E[X|P_{S_{Y|X}}X] = P_{S_{Y|X}}X$, and SAVE and pHd require an additional condition called the constant variance condition, which is $\mathrm{cov}[X|P_{S_{Y|X}}X] = Q_{S_{Y|X}}$, where $Q_{S_{Y|X}} = I_p - P_{S_{Y|X}}$. These conditions are satisfied when the distribution of X is elliptically contoured or multivariate normal. For a detailed discussion of the conditions, readers can consult Cook (1998).

Although SIR, SAVE, pHd and IHT discussed above work well in practice, there has not been much study regarding when they can exhaustively recover the central subspace or the central mean subspace in the literature. It is known that SIR fails to capture directions along which Y is symmetric about. For example, if $Y = (\beta^\tau X)^2 + \varepsilon$ where β is a p-dimensional vector, τ denotes transpose and ε is a random error independent of X, SIR will miss β. A sufficient condition for SAVE to exhaustively recover $S_{Y|X}$ is that the conditional distribution of X given Y is multivariate normal, which may be restrictive in practice. A potential risk of applying these methods is that they may lead to the loss of information regarding $F_{Y|X}$ or $E[Y|X]$. Therefore, it is desirable to derive new methods that can guarantee the exhaustive recovery of the central subspace or the central mean subspace under general conditions. Recently, Li, Zha and Chiaromonte (2005) made progress in this direction by proposing *contour regression* for sufficient dimension reduction. Contour regression assumes that X follows an elliptically-contoured distribution, and in addition, it requires an intriguing condition that involves X, Y and the vectors in $S_{Y|X}$ and $S_{Y|X}^{\perp}$ (Assumption 2.1, Li Zha and Chiaromonte 2005), where $S_{Y|X}^{\perp}$ is the complementary subspace of $S_{Y|X}$. The latter assumption is reasonable, but it serves primarily as a technical assumption and may not be easily verified in practice.

The current paper represents another effort to derive methods that can fully recover the central mean subspace and the central subspace under various conditions. The primary tool we use is the Fourier transform. At the population level, we have derived two candidate matrices M_{FM} and M_{FC} whose column spaces are identical to the central mean subspace and the central subspace, respectively.

Given a sample of (X, Y), if consistent estimates of M_{FM} and M_{FC} can be found, they can be used to exhaustively recover the central mean subspace and the central subspace. In fact, the consistent estimates exist, but their exact formulas or calculations depend on the amount of prior knowledge we have regarding the distribution of X. Due to limited space, in this paper, we only fully implement our methods for the case where X is normally distributed. The implementation of our method under more general conditions on X is briefly described in this paper and is currently under further investigation, and we will report the results elsewhere in the future.

To facilitate our approach, we need to modify the model assumptions as follows. First we assume that the joint distribution of (X, Y), the conditional distributions of $X|Y$ and $Y|X$, and the marginal distributions of X and Y admit densities, which are denoted by $f_{X,Y}(x, y)$, $f_{Y|X}(y|x)$, $f_{X|Y}(x|y)$, $f_X(x)$ and $f_Y(y)$, respectively. Let $B = (\beta_1, \beta_2, \ldots, \beta_q)$ be a $p \times q$ matrix with its columns forming a basis for $\mathcal{S}_{Y|X}$. Then (1) can be restated in terms of conditional distributions, that is, $F_{Y|X} = F_{Y|B^\tau X}$, or in terms of density,

$$f_{Y|X}(y|x) = f_{Y|B^\tau X}(y|B^\tau x) = h(y; \beta_1^\tau x, \beta_2^\tau x, \ldots, \beta_q^\tau x), \tag{3}$$

where $h(y; u_1, \ldots, u_q)$ is a $(q+1)$-variate function. Similarly, let $\alpha_1, \ldots, \alpha_q$ be a basis of $S_{E[Y|X]}$, then (2) is equivalent to

$$E[Y \mid X = x] = g(\alpha_1^\tau x, \alpha_2^\tau x, \ldots, \alpha_q^\tau x), \tag{4}$$

where g is a q-variate function. We assume the differentiability of h and g with respect to their coordinates wherever it is needed.

The rest of the paper is organized as follows. Section 2 derives M_{FM} for central mean subspace and Section 3 derives M_{FC} for central subspace. Section 4 derives the estimates of M_{FM} and M_{FC} under the assumption that X is normal, and discusses the asymptotic properties of these estimates. Section 5 focuses on the implementation of the proposed methods for estimating central subspace and central mean subspace. In Section 6, the proposed methods are compared to SIR, SAVE and other existing methods using synthetic and real data. Section 7 contains conclusions and future work. The proofs of propositions, theorems and crucial equations are given in Appendix 8. In this paper, we use $\mathcal{S}(M)$ to denote the linear space spanned by the columns of a matrix M.

2 Central Mean Subspace

In this section, we propose a candidate matrix M_{FM} whose column space is exactly the central mean subspace $\mathcal{S}_{E[Y|X]}$. We follow a commonly used idea for deriving candidate matrices in the literature,

which is to identify vectors that belong to $\mathcal{S}_{E[Y|X]}$ and combine them to generate the candidate matrix. The major tool we use is the Fourier transform.

Let $m(x) = E[Y|X = x]$. From (4), $m(x) = g(u)$ where $u = A^\tau x$ and $A = (\alpha_1, \alpha_2, \ldots, \alpha_q)$ whose columns form a basis of $\mathcal{S}_{E[Y|X]}$. Let $\frac{\partial}{\partial x} = (\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \ldots, \frac{\partial}{\partial x_p})^\tau$ denote the gradient operator. By the chain rule of differentiation,

$$\frac{\partial}{\partial x} m(x) = A \, \frac{\partial}{\partial u} g(u).$$

Thus the gradient of $m(x)$ at any fixed x is a linear combination of $\alpha_1, \alpha_2, \ldots, \alpha_q$, therefore it is in $\mathcal{S}_{E[Y|X]}$. Let $\mathrm{supp}(X) = \{x \in \mathbb{R}^p : f_X(x) > 0\}$ be the support of X. The collection of all the gradients of $m(x)$ over $x \in \mathrm{supp}(X)$ spans the central mean subspace $\mathcal{S}_{E[Y|X]}$, so does the collection of all the gradients of $m(x)$ weighted by $f_X(x)$, that is,

$$\mathcal{S}_{E[Y|X]} = \mathrm{span}\{\tfrac{\partial}{\partial x} m(x), x \in \mathrm{supp}(X)\} = \mathrm{span}\{(\tfrac{\partial}{\partial x} m(x))\, f_X(x), x \in \mathbb{R}^p\}. \tag{5}$$

The proof of (5) is given in Appendix 8. From (5), it appears that we can immediately use the gradient of $m(x)$ to generate a candidate matrix for $\mathcal{S}_{E[Y|X]}$. However, this idea leads to inefficient estimation of the central mean subspace, because much effort has to be spent on estimating g and its derivatives nonparametrically (Hristache et al. 2001; Xia et al. 2002). Recall that the primary goal of sufficient dimension reduction is to recover the central mean subspace $\mathcal{S}_{E[Y|X]}$ only. So we hope to achieve dimension reduction while avoiding fitting the link function $g(x)$ and its derivatives as much as possible. This can be realized by considering the Fourier transform of the gradient of $m(x)$ instead of using $\frac{\partial}{\partial x} m(x)$ directly. Another advantage of using the Fourier transform of $\frac{\partial}{\partial x} m(x)$ is that sufficient dimension reduction for the central subspace and the central mean subspace can be dealt with in a unified fashion as will be demonstrated in the next section.

For $\omega \in \mathbb{R}^p$, let

$$\psi(\omega) = \int \exp\{\imath\, \omega^\tau x\} \left(\tfrac{\partial}{\partial x}\, m(x)\right) f_X(x)\, dx. \tag{6}$$

Then $\psi(\omega)$ is the Fourier transform of the density-weighted gradient $(\frac{\partial}{\partial x}\, m(x)) f_X(x)$. Intuitively, $\psi(\omega)$ can also be regarded as an average of the gradient of $m(x)$ weighted by $\exp\{\imath\, \omega^\tau x\}$ over x with density $f_X(x)$. In particular, when $\omega = 0$, $\psi(0) = E[\frac{\partial}{\partial x} m(X)]$, which is exactly the average gradient of $m(x)$ (Härdle and Stoker 1989). Therefore, in some sense, $\psi(\omega)$ is a generalized average derivative of the mean response $m(x)$. Let $a(\omega)$ and $b(\omega)$ be the real and imaginary parts of $\psi(\omega)$, that is, $\psi(\omega) = a(\omega) + \imath b(\omega)$. Because the gradient of $m(x)$ belongs to $\mathcal{S}_{E[Y|X]}$, it implies that both $a(\omega)$ and $b(\omega)$ belong to $\mathcal{S}_{E[Y|X]}$.

An appealing property of $\psi(\omega)$ is that it contains all the information of the gradient $\frac{\partial}{\partial x}m(x)$, because $\frac{\partial}{\partial x}m(x)$ can be recovered from $\psi(\omega)$ through the inverse Fourier transform (Folland 1992; Page 244). Assuming $(\frac{\partial}{\partial x}m(x))f_X(x)$ is integrable and continuous on $\mathbb{R}^p$ and $\psi(\omega)$ is also integrable,

$$(\frac{\partial}{\partial x}m(x))f_X(x) = (2\pi)^{-p}\int \exp\{-\imath x^\tau\omega\}\,\psi(\omega)\,d\omega. \tag{7}$$

From (5), we know that the central mean subspace is spanned by the gradients. Considering the correspondence between $\psi(\omega)$ and $\frac{\partial}{\partial x}m(x)$ as demonstrated in (6) and (7), we can use $\psi(\omega)$ to generate a candidate matrix for the central mean subspace $\mathcal{S}_{E[Y|X]}$. The other properties of $\psi(\omega)$ are summarized in the following proposition.

Proposition 1.

1. *Both $a(\omega)$ and $b(\omega)$ are vectors in $\mathcal{S}_{E[Y|X]}$, furthermore, $\mathcal{S}_{E[Y|X]} = \text{span}\{a(\omega), b(\omega) : \omega \in \mathbb{R}^p\}$.*

2. *Suppose $\log f_X(x)$ is differentiable and $m(x)f_X(x)$ goes to zero as $\|x\| \to \infty$, then*

$$\psi(\omega) = -E_{(X,Y)}\big[Y(\imath\omega + G(X))\,\exp\{\imath\,\omega^\tau X\}\big], \tag{8}$$

   where $G(x) = \frac{\partial}{\partial x}\log f_X(x)$.

3. *If $(\frac{\partial}{\partial x_i}m(x))f_X(x)$ is absolutely integrable for $1 \le i \le p$, then $\psi(\omega) \to 0$ as $\|\omega\| \to \infty$.*

4. *If $(\frac{\partial}{\partial x_i}m(x))f_X(x)$ is squared integrable for $1 \le i \le p$, then*

$$\int (\tfrac{\partial}{\partial x}m(x))(\tfrac{\partial}{\partial x}m(x))^\tau f_X(x)^2\,dx = (2\pi)^{-p}\int \psi(\omega)\bar{\psi}(\omega)^\tau\,d\omega, \tag{9}$$

   where $\bar{\psi}(\omega)$ is the conjugate of $\psi(\omega)$.

The first property indicates that the real and imaginary parts of $\psi(\omega)$ are the vectors we can use to generate candidate matrices for the central mean subspace. In the second property, $\psi(\omega)$ is represented as an expectation of the random function $-Y(\imath\omega + G(X))\exp\{\imath\,\omega^\tau X\}$. Recall that $\psi(\omega)$ was originally defined in terms of $m(x)$ as in (6). The new expression of $\psi(\omega)$ in (8) does not include $m(x)$ explicitly, which provides us the opportunity to estimate $\psi(\omega)$ without directly estimating $m(x)$. This distinguishes our method from the methods that directly estimate $m(x)$ and its derivatives in the literature.

The third property is essentially the Riemann-Lebesgue Lemma for the Fourier transform (Folland 1992; Pages 217, 243). It indicates that, when the density-weighted gradient is absolutely integrable, $\psi(\omega)$ decays to zero as the norm of ω goes to infinity. The fourth property is a result from applying the

Plancheral Theorem for the Fourier transform (Folland 1992; Pages 222, 224) to $(\frac{\partial}{\partial x}m(x))f_X(x)$ and $\psi(\omega)$, and it establishes the connection between the expected outer product of the density-weighted gradient of $m(x)$, that is, $E\big[(\frac{\partial}{\partial x}m(X))(\frac{\partial}{\partial x}m(X))^\tau f_X(X)\big]$, and the integral of the outer-product of $\psi(\omega)$. Let $M^*_{\text{FM}} = (2\pi)^{-p}\int \psi(\omega)\bar{\psi}(\omega)^\tau d\omega$. Then the column space of M^*_{FM}, $\mathcal{S}(M^*_{\text{FM}})$, is exactly equal to the central mean subspace as stated in the following proposition.

Proposition 2. *M^*_{FM} is a real non-negative definite matrix and $\mathcal{S}(M^*_{FM}) = \mathcal{S}_{E[Y|X]}$.*

Intuitively, M^*_{FM} can be considered as the sum of the outer product of $\psi(\omega)$ over all ω. Because of (7), $\psi(\omega)$ can be regarded as the vector of coefficients of $\exp\{\imath x^\tau\omega\}$ in the representation of $\frac{\partial}{\partial x}m(x)f_X(x)$. For different ω's, $\exp\{\imath x^\tau\omega\}$ are basic oscillatory functions with different frequencies. So $\psi(\omega)$ with small $\|\omega\|$ captures the patterns of $\frac{\partial}{\partial x}m(x)f_X(x)$ with low frequencies, while $\psi(\omega)$ with large $\|\omega\|$ captures the patterns of $\frac{\partial}{\partial x}m(x)f_X(x)$ with high frequencies. According to the third property of $\psi(\omega)$, $\psi(\omega)$ goes to zero when $\|\omega\|$ goes to infinity. Therefore, when patterns with various frequencies are of different interests, $\psi(\omega)$ of different ω should be treated differently. This can be realized by assigning different weights to ω when combining the outer product of $\psi(\omega)$. We use $K(\omega)$ to denote the weight function, and generate a more flexible candidate matrix for the central mean subspace as follows,

$$M_{\text{FM}} = \text{Re}\left(\int \psi(\omega)\bar{\psi}(\omega)^\tau K(\omega)\, d\omega\right) = \int [a(\omega)a(\omega)^\tau + b(\omega)b(\omega)^\tau]K(\omega)\, d\omega, \tag{10}$$

where Re() means "the real part of". Note if $K(\omega)$ is a radial weight function, then $\int \psi(\omega)\bar{\psi}(\omega)^\tau K(\omega)d\omega$ itself is a real matrix.

Proposition 3. *If $K(\omega)$ is a positive weight function on $\mathbb{R}^p$, then M_{FM} is a non-negative definite matrix and $\mathcal{S}(M_{FM}) = \mathcal{S}_{E[Y|X]}$.*

Proposition 3 indicates that the central mean subspace $\mathcal{S}_{E[Y|X]}$ can be exhaustively recovered by the column space of M_{FM}. In the proposition, we have assumed that $K(\omega)$ is positive over all $\mathbb{R}^p$. This condition can be substantially weakened. For example, if $\psi(\omega)$ is analytic, then the proposition holds true for any weight function $K(\omega)$ with bounded support that contains an open set. In this paper, we will focus on positive weight functions only. Though the proposition is true for any positive function in theory, the particular choice will affect the performance of dimension reduction based on M_{FM} in practice, especially when the sample size is moderate. For simplicity, we choose the Gaussian function $K(\omega) = (2\pi)^{-p/2}\exp\{-\|\omega\|^2/2\sigma^2_{\text{W}}\}$ as the weight function in the rest of the paper. Note that σ^2_{W} in $K(\omega)$ is a constant (or tuning parameter) that controls how the weight is assigned to different $\psi(\omega)$'s.

Furthermore, $K(\omega)$ leads to an explicit expression of M_{FM} given below,

$$M_{\text{FM}} = E_{(U_1,V_1),(U_2,V_2)}\left[J_{\text{FM}}((U_1,V_1),(U_2,V_2))\right], \tag{11}$$

where

$$J_{\text{FM}}((U_1,V_1),(U_2,V_2)) = V_1V_2\, e^{-\frac{\sigma_{\text{W}}^2}{2}\|U_{12}\|^2}\left[\sigma_{\text{W}}^2 I_p + (G(U_1)-\sigma_{\text{W}}^2U_{12})(G(U_2)+\sigma_{\text{W}}^2U_{12})^\tau\right],$$

(U_1,V_1) and (U_2,V_2) are independent and identically distributed as (X,Y), I_p is the $p\times p$ identity matrix, and $U_{12}=U_1-U_2$.

3 Central Subspace

As discussed in the introduction, the central mean subspace can only capture the information in X regarding the mean response $E[Y|X]$. In applications where the entire conditional distribution $F_{Y|X}$ is of interest, sufficient dimension reduction should aim at the central subspace $\mathcal{S}_{Y|X}$. This section is focused on the derivation of the candidate matrix M_{FC} for $\mathcal{S}_{Y|X}$. We will again use the Fourier transform as well as other similar ideas from Section 2.

First, we establish a connection between $\mathcal{S}_{Y|X}$ and a family of central mean subspaces. As noted in the introduction, the central mean subspace $\mathcal{S}_{E[Y|X]}$ is always a subspace of $\mathcal{S}_{Y|X}$. Let T denote a transformation of Y. Then $T(Y)$ is a new response. It can be shown that the central mean subspace of $T(Y)$, denoted by $\mathcal{S}_{E[T(Y)|X]}$, is also a subspace of $\mathcal{S}_{Y|X}$. For two different transformations $T_1(Y)$ and $T_2(Y)$, their corresponding central mean subspaces $\mathcal{S}_{E[T_1(Y)|X]}$ and $\mathcal{S}_{E[T_2(Y)|X]}$ are not necessarily identical and may cover different part of $\mathcal{S}_{Y|X}$. This provides a possibility to recover the entire central subspace by collecting the central mean subspace of $T(Y)$ over all the possible transformations, that is,

$$\mathcal{S}_{Y|X} = \sum_{\text{all possible } T} \mathcal{S}_{E[T(Y)|X]}. \tag{12}$$

The above equation is indeed true and will be implied by Proposition 4 given below. In fact, it is not necessary to use all the possible transformations in (12), a family of properly chosen transformations is enough to serve the purpose.

For any given $t\in\mathbb{R}$, define $T(Y,t)=\exp\{\imath tY\}$. The $T(\cdot,t)$'s form a family of transformations indexed by t. The mean response of $T(Y,t)$ at $X=x$ is

$$m(x,t) = E[T(Y,t)|X=x] = \int \exp\{\imath ty\}\, f_{Y|X}(y\mid x)\, dy.$$

So $m(x,t)$ is the Fourier transform, or the characteristic function, of the conditional density function $f_{Y|X}(y|x)$. Note that both $T(Y,t)$ and $m(x,t)$ are complex functions. The central mean subspace for $T(Y,t)$ is defined to be the sum of the central mean subspaces for its real and imaginary parts, that is, $\mathcal{S}_{E[T(Y,t)|X]} = \mathcal{S}_{E[\sin(tY)|X]} + \mathcal{S}_{E[\cos(tY)|X]}$. The following proposition states that the central subspace $\mathcal{S}_{Y|X}$ is equal to the sum of $\mathcal{S}_{E[T(Y,t)|X]}$ over $t \in \mathbb{R}$.

Proposition 4. *Suppose $\frac{\partial}{\partial x} f_{Y|X}(y|x)$ exists and is absolutely integrable with respect to y. Then*

$$\mathcal{S}_{Y|X} = \sum_{t\in\mathbb{R}} \mathcal{S}_{E[T(Y,t)|X]}.$$

When defining the central kth moment subspace, Yin and Cook (2002) considered the power transformations of Y, which are Y^l with $1 \le l \le k$. The transformation we consider can be regarded as an extension of the power transformation, they, however, lead to entirely different methods for recovering the central subspace. Proposition 4 suggests that, in order to recover the central subspace $\mathcal{S}_{Y|X}$, we can first recover the central mean subspace for $T(Y,t)$ at fixed t, then combine them over $t \in \mathbb{R}$. The methods and results developed in Section 2 for the central mean subspace $\mathcal{S}_{E[Y|X]}$ can be directly generalized for the central mean subspace $\mathcal{S}_{E[T(Y,t)|X]}$ with Y replaced by $\exp\{\imath tY\}$. Hence, we can combine the candidate matrices for $\mathcal{S}_{E[T(Y,t)|X]}$ to generate a candidate matrix for $\mathcal{S}_{Y|X}$. The idea of combining candidate matrices to generate new ones was originally mentioned in the rejoinder of Li (1991) and it was further investigated in Ye and Weiss (2003). In the following, we start with applying the Fourier transform to the gradient of $m(x,t)$ as in Section 2 and materialize the idea outlined above to derive a candidate matrix M_{FC} for $\mathcal{S}_{Y|X}$.

Because $m(x,t)$ is the mean response of $T(Y,t)$ at $X = x$, similar to (5) in Section 2, we have

$$\mathcal{S}_{E[T(Y,t)|X]} = \text{span}\{\tfrac{\partial}{\partial x} m(x,t), x \in \text{supp}(X)\} = \text{span}\{(\tfrac{\partial}{\partial x} m(x,t))\, f_X(x), x \in \mathbb{R}^p\}.$$

Note that the spanned space of complex vectors is defined to be the space spanned by their real parts and imaginary parts. Using Proposition 4, we have

$$\mathcal{S}_{Y|X} = \text{span}\{\tfrac{\partial}{\partial x} m(x,t), x \in \text{supp}(X), t \in \mathbb{R}\} = \text{span}\{(\tfrac{\partial}{\partial x} m(x,t))\, f_X(x), x \in \mathbb{R}^p, t \in \mathbb{R}\}. \tag{13}$$

As in Section 2, to derive a candidate matrix for $\mathcal{S}_{Y|X}$, we do not use the gradient $\frac{\partial}{\partial x} m(x,t)$ directly, instead we consider its Fourier transform. For any $\omega \in \mathbb{R}^p$ and $t \in \mathbb{R}$, define

$$\phi(\omega, t) = \int \exp\{\imath \omega^\tau x\}\, (\tfrac{\partial}{\partial x} m(x,t))\, f_X(x)\, dx. \tag{14}$$

Then $\phi(\omega, t)$ is the Fourier transform of $\frac{\partial}{\partial x} m(x,t)$ weighted by the marginal density function of X, and it preserves all the information about $\frac{\partial}{\partial x} m(x,t)$. Let $a(\omega,t)$ and $b(\omega,t)$ be the real and imaginary parts of $\phi(\omega,t)$. The properties of $\phi(\omega,t)$ are summarized in the following proposition.

Proposition 5.

1. *Both $a(\omega,t)$ and $b(\omega,t)$ are vectors in $\mathcal{S}_{E[T(Y,t)|X]}$, furthermore,*

$$\mathcal{S}_{Y|X} = \text{span}\{a(\omega,t), b(\omega,t) : \omega \in \mathbb{R}^p, t \in \mathbb{R}\}. \tag{15}$$

2. *Suppose $\log f_X(x)$ is differentiable and $m(x,t)\log f_X(x)$ goes to zero when $\|x\| \longrightarrow \infty$,*

$$\phi(\omega,t) = -E_{(X,Y)}[(\imath\omega + G(X))\, \exp\{\imath\omega^\tau X + \imath t Y\}]. \tag{16}$$

3. *If $(\frac{\partial}{\partial x_i} m(x,t)) f_X(x)$ is absolutely integrable for $1 \le i \le p$, then $\phi(\omega,t) \to 0$ as $\|\omega\| \to \infty$.*

4. *If $(\frac{\partial}{\partial x_i} m(x,t)) f_X(x)$ is squared integrable for $1 \le i \le p$, then*

$$\int (\tfrac{\partial}{\partial x} m(x,t))(\tfrac{\partial}{\partial x} m(x,t))^\tau f_X(x)^2\, dx = (2\pi)^{-p} \int \phi(\omega,t)\bar{\phi}(\omega,t)^\tau\, d\omega, \tag{17}$$

where $\bar{\phi}(\omega,t)$ is the conjugate of $\phi(\omega,t)$.

Proposition 5 is a direct extension of Proposition 1 with Y replaced by $\exp\{\imath t Y\}$ and $m(x)$ replaced by $m(x,t)$. According to the first property in Proposition 5, using the similar arguments as following Proposition 1, we can combine $a(\omega,t)$ and $b(\omega,t)$ to derive a candidate matrix for $\mathcal{S}_{Y|X}$. Let $K(\omega)$ be a weight function for $\omega \in \mathbb{R}^p$ and $k(t)$ be a weight function for $t \in \mathbb{R}$. Define

$$M_{\text{FC}} = \text{Re}\left(\iint \phi(\omega,t)\bar{\phi}(\omega,t)^\tau K(\omega)k(t)\, d\omega dt\right) = \iint [a(\omega,t)a(\omega,t)^\tau + b(\omega,t)b(\omega,t)^\tau] K(\omega)k(t)\, d\omega dt. \tag{18}$$

Proposition 6. *If $K(\omega)$ is a positive weight function on $\mathbb{R}^p$ and $k(t)$ is a positive weight function on $\mathbb{R}$, then M_{FC} is a non-negative definite matrix and $\mathcal{S}(M_{FC}) = \mathcal{S}_{Y|X}$.*

Proposition 6 implies that $\mathcal{S}_{Y|X}$ can be exhaustively recovered by the column space of M_{FC}. In this paper, for convenience, both $K(\omega)$ and $k(t)$ are chosen to be the Gaussian functions, which are $K(\omega) = (2\pi)^{-p/2}\exp\{-\|\omega\|^2/2\sigma_{\text{W}}^2\}$ and $k(t) = (2\pi)^{-1/2}\exp\{-t^2/2\sigma_{\text{T}}^2\}$, where σ_{W}^2 and σ_{T}^2 are two constants that control how the weights are assigned to different $\phi(\omega,t)$. Furthermore, the Gaussian weight functions lead to an explicit expression of M_{FC} given below,

$$M_{\text{FC}} = E_{(U_1,V_1),(U_2,V_2)}\left[J_{\text{FC}}((U_1,V_1),(U_2,V_2))\right], \tag{19}$$

where

$$J_{\mathrm{FC}}((U_1,V_1),(U_2,V_2)) = \exp\{-\frac{\sigma_{\mathrm{W}}^2}{2}\|U_{12}\|^2 - \frac{\sigma_{\mathrm{T}}^2}{2}V_{12}^2\}\left[\sigma_{\mathrm{W}}^2 I_p + (G(U_1) - \sigma_{\mathrm{W}}^2 U_{12})(G(U_2) + \sigma_{\mathrm{W}}^2 U_{12})^\tau\right], \tag{20}$$

(U_1, V_1) and (U_2, V_2) are independent and identically distributed as (X, Y), I_p is the identity matrix, $U_{12} = U_1 - U_2$, and $V_{12} = V_1 - V_2$.

The derivation of M_{FC} can also be understood from the perspective of inverse regression used in SIR and SAVE, where the candidate matrices were defined by the first and second moments of the conditional distribution of X given Y. Next we first present a connection between $\phi(\omega, t)$, which is used to define M_{FC}, and the conditional distribution of X given Y. Define

$$\eta(y,\omega) = -E[(\imath\omega + G(X))\exp\{\imath\omega^\tau x\} \mid Y = y].$$

For fixed ω, $\eta(y, \omega)$ can be regarded as the mean response vector for inversely regressing $-(\imath\omega + G(X))\exp\{\imath\omega^\tau X\}$ on Y. The properties of $\eta(y, \omega)$ are given in the following proposition.

Proposition 7. *Suppose for any fixed y, $f_{X,Y}(x, y)$ goes to zero as $\|x\| \longrightarrow \infty$.*

1. *For any given y and ω, the real and imaginary parts of $\eta(y, \omega)$ are vectors in $\mathcal{S}_{Y|X}$, in particular, those of $\eta(y, 0) = -E[G(X)|Y = y]$ are vectors in $\mathcal{S}_{Y|X}$.*
2. *$\phi(\omega, t) = E[\eta(Y, \omega)\exp\{\imath tY\}]$.*

Recall that $\phi(\omega, t)$ serves as the building blocks for M_{FC}. From the second statement in Proposition 7, $\phi(\omega, t)$ can be regarded as the Fourier transform of $\eta(y, \omega)$ with respect to the marginal density of Y. When X is multivariate normal with mean zero and covariance matrix I_p, $G(x) = -x$ and $\eta(y, 0) = E[X|Y]$, which serves as the building blocks in SIR. This observation leads to an important connection between M_{FC} and the candidate matrix M_{SIR} used in SIR.

Proposition 8. *Suppose X follows the normal distribution with mean 0 and covariance matrix I_p. If $K(\omega)$ is chosen to be a point mass at $\omega = 0$, then $\mathcal{S}(M_{FC}) = \mathcal{S}(M_{SIR})$.*

The above proposition indicates that the column spaces of M_{FC} and M_{SIR} coincide when X follows the standard normal distribution and the weight function $K(\omega)$ is degenerate at $\omega = 0$. Hence, M_{SIR} could be considered as a special case of M_{FC} under the normality assumption. Note that, although $\mathcal{S}(M_{\mathrm{FC}})$ and $\mathcal{S}(M_{\mathrm{SIR}})$ are the same, the matrices are generally different from each other. It is known that SIR fails to capture directions along which Y is symmetric about, so does M_{FC} with $\sigma_{\mathrm{W}}^2 = 0$. In general, we will not use a degenerate weight function for ω. When $\sigma_{\mathrm{W}}^2 > 0$, the entire central subspace can be successfully recovered as stated in Proposition 6.

4 Estimation of Candidate Matrices

In this section, we derive the estimates of M_{FC} and M_{FM} and discuss their asymptotic properties. We assume that the dimensionality q and the tuning parameters σ_{W}^2 and σ_{T}^2 are known and will discuss their selection in the next section.

Let $\{(x_i, y_i)\}_{1\le i\le n}$ be a random sample of (X, Y). Without loss of generality, we assume $E[X] = 0$ and $cov[X] = I_p$. Let us consider M_{FC} first. According to (19), M_{FC} is the expectation of J_{FC} over (U_1, V_1) and (U_2, V_2) that are independent and identically distributed as (X, Y). Let $F(x, y)$ be the cumulative distribution function of (X, Y). Then M_{FC} can be expressed as

$$M_{\text{FC}} = \iint J_{\text{FC}}((u_1, v_1), (u_2, v_2))\, dF(u_1, v_1) dF(u_2, v_2). \tag{21}$$

With the given sample $\{(x_i, y_i)\}_{1\le i\le n}$, a natural estimate of $F(x, y)$ is its empirical distribution,

$$F_n(x, y) = \frac{1}{n}\sum_{i=1}^{n} I_{[x_i\le x,\, y_i\le y]},$$

where $I_{[\cdot]}$ is the indicator function. Therefore, a proper estimate of M_{FC} is derived by replacing $F(u_1, v_1)$ and $F(u_2, v_2)$ in (21) with $F_n(u_1, v_1)$ and $F_n(u_2, v_2)$ respectively, which is

$$\hat{M}_{\text{FC}} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} J_{\text{FC}}((x_i, y_i), (x_j, y_j)). \tag{22}$$

The explicit expression of $J_{\text{FC}}((x_i, y_i), (x_j, y_j))$ is given in (20). Notice that, to make $\hat{M}_{\text{FC}}$ a legitimate estimate, we need to estimate $G(x_i) = \frac{\partial}{\partial x}\log f_X(x_i)$ for $1 \le i \le n$.

If we know that the distribution of X belongs to a certain parametric family, that is, $f_X(x) = f_0(x; \theta)$, where $f_0(\cdot)$ is of known form and θ is an unknown parameter, then the maximum likelihood estimate $\hat{\theta}$ can be calculated using $\{x_i\}_{1\le i\le n}$, and $G(x_i)$ can be estimated by $\frac{\partial}{\partial x} f_0(x_i; \hat{\theta})/f_0(x_i; \hat{\theta})$. If $f_X(x)$ belongs to the family of elliptically contoured distributions, that is, $f_X(x) = g(\|x\|^2)$, where $g(\cdot)$ is an unknown function, then using $\{\|x_i\|\}_{1\le i\le n}$, a one-dimensional nonparametric procedure can be employed to obtain $\hat{g}(\|x\|^2)$ and $\hat{g}'(\|x\|^2)$, which are the estimates of g and its derivative g', and $\frac{\partial}{\partial x}\log(g(\|x\|^2))$ can be estimated by $2x\hat{g}^{-1}(\|x\|^2)\hat{g}'(\|x\|^2)$. In general, if there is no prior knowledge about $f_X(x)$, nonparametric density estimators can be used to estimate $f_X(x)$ and $\frac{\partial}{\partial x} f_X(x)$, and further to obtain an estimate of $G(x_i)$. The general case is currently under investigation.

In the rest of this paper, we will focus on the case where X follows a multivariate normal distribution only. Normality is a common assumption in regression and is at least approximately valid in many applications. For applications where the normality assumption is not valid, variable transformation or data resampling can be considered to alleviate the violation of normality so that the methods

developed below can still be applied; see Brillinger (1991) for the resampling method and Cook and Nachtsheim (1994) for the Voronoi weighting method.

Assume X follows the multivariate normal distribution with mean 0 and covariance matrix I_p. Then $G(x) = -x$. For clarity, we use J_{FCn}, M_{FCn} and $\hat{M}_{\mathrm{FCn}}$ to denote J_{FC}, M_{FC} and $\hat{M}_{\mathrm{FC}}$ under the normality assumption, respectively. Hence,

$$M_{\mathrm{FCn}} = E_{(U_1,V_1),(U_2,V_2)}\left[J_{\mathrm{FCn}}((U_1,V_1),(U_2,V_2))\right], \tag{23}$$

and

$$\hat{M}_{\mathrm{FCn}} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} J_{\mathrm{FCn}}((u_i,v_i),(u_j,v_j)), \tag{24}$$

where

$$J_{\mathrm{FCn}}((u_1,v_1),(u_2,v_2)) = \exp\{-\frac{\sigma_{\mathrm{W}}^2}{2}\|u_{12}\|^2 - \frac{\sigma_{\mathrm{T}}^2}{2}v_{12}^2\}\left[\sigma_{\mathrm{W}}^2 I_p + (u_1+\sigma_{\mathrm{W}}^2 u_{12})(u_2-\sigma_{\mathrm{W}}^2 u_{12})^\tau\right]. \tag{25}$$

Note $\hat{M}_{\mathrm{FCn}}$ is a V-statistic and it can be expanded as the sum of a U-statistic and a low-order term (Lee 1990). The asymptotic distribution of $\hat{M}_{\mathrm{FCn}}$ can be obtained by the theory of U-statistic.

Theorem 1. *Suppose X follows the standard multivariate normal distribution and the covariance matrix of $vec(J_{FCn}((U_1,V_1),(U_2,V_2)))$ exists. As $n\to\infty$,*

$$\hat{M}_{FCn} = M_{FCn} + \frac{1}{n}\sum_{i=1}^{n}(J_{FCn}^{(1)}(x_i,y_i) - 2M_{FCn}) + o_p(n^{-1/2}),$$

where

$$J_{FCn}^{(1)}(x,y) = E_{(U_2,V_2)}[J_{FCn}((x,y),(U_2,V_2)) + J_{FCn}((x,y),(U_2,V_2))^\tau].$$

Let Σ_{FCn} be the covariance matrix of $vec(J_{FCn}^{(1)}(X,Y))$. Then

$$\sqrt{n}(vec(\hat{M}_{FCn}) - vec(M_{FCn})) \xrightarrow{\mathcal{L}} N(0,\Sigma_{FCn}), \quad as\ n\to\infty.$$

In the theorem above, *vec* is an operator that transforms a matrix to a vector by stacking up all its columns. For instance, if $M = (m_1,\ldots,m_k)$ is a $p\times k$ matrix and m_i's are the column vectors, then $vec(M) = (m_1^\tau,\ldots,m_k^\tau)^\tau$ is a $kp\times 1$ vector. Theorem 1 asserts $\hat{M}_{\mathrm{FCn}}$ converges to M_{FCn} at the rate of $\sqrt{n}$, which implies that the eigenvalues and eigenvectors of $\hat{M}_{\mathrm{FCn}}$ converges to those of M_{FCn} at the same rate.

Although the explicit expression of M_{FCn} in (23) is obtained under the normality assumption, in fact, it remains to be a candidate matrix for $\mathcal{S}_{Y|X}$ under a weaker condition as stated in the following proposition.

Proposition 9. *Suppose $B = (\beta_1, \ldots, \beta_q)$ is an orthonormal basis of $\mathcal{S}_{Y|X}$ and $\tilde{B} = (\beta_{q+1}, \ldots, \beta_p)$ is an orthonormal basis of the complementary space of $\mathcal{S}_{Y|X}$ in $\mathbb{R}^p$. If $B^\tau X$ and $\tilde{B}^\tau X$ are independent with each other and $\tilde{B}^\tau X$ follows the standard normal distribution, then $\mathcal{S}(M_{FCn}) \subseteq \mathcal{S}_{Y|X}$.*

We call the assumption in Proposition 9 as the weak normality condition. Proposition 9 implies that M_{FCn} can be used to recover the central subspace under the weak normality assumption, though it does not guarantee that the central subspace can be recovered exhaustively as under the normality condition. Note that the distribution of $B^\tau X$ can be arbitrary. The weak normality condition represents a situation in practice where Y only depends on a subset of predictors and the rest of the predictors are random noises following normal distributions.

For the central mean subspace $\mathcal{S}_{E[Y|X]}$ and its candidate matrix M_{FM}, similar discussions can be used to derive the estimates of M_{FM} under various conditions on X. In the following, we only report the results under the normality assumption on X. Again, we use J_{FMn}, M_{FMn} and $\hat{M}_{\mathrm{FMn}}$ for J_{FM}, M_{FM} and $\hat{M}_{\mathrm{FM}}$ under the normality condition. Recall that $G(x) = -x$. Therefore,

$$M_{\mathrm{FMn}} = E_{(U_1,V_1),(U_2,V_2)}\left[J_{\mathrm{FMn}}((U_1,V_1),(U_2,V_2))\right], \tag{26}$$

and the estimate of M_{FMn} is

$$\hat{M}_{\mathrm{FMn}} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} J_{\mathrm{FMn}}((x_i,y_i),(x_j,y_j)), \tag{27}$$

where

$$J_{\mathrm{FMn}}((u_1,v_1),(u_2,v_2)) = v_1 v_2 \exp\{-\frac{\sigma_{\mathrm{W}}^2}{2}\|u_{12}\|^2\}\left[\sigma_{\mathrm{W}}^2 I_p + (u_1 + \sigma_{\mathrm{W}}^2 u_{12})(u_2 - \sigma_{\mathrm{W}}^2 u_{12})^\tau\right]. \tag{28}$$

The asymptotic normality of $\hat{M}_{\mathrm{FMn}}$ is established in the following theorem.

Theorem 2. *Suppose X follows the standard multivariate normal distribution and the covariance matrix of $vec(J_{FMn}((U_1,V_1),(U_2,V_2)))$ exists. As $n \to \infty$,*

$$\hat{M}_{FMn} = M_{FMn} + \frac{1}{n}\sum_{i=1}^{n}(J_{FMn}^{(1)}(x_i,y_i) - 2M_{FMn}) + o_p(n^{-1/2}),$$

where

$$J_{FMn}^{(1)}(x,y) = E_{(U_2,V_2)}[J_{FMn}((x,y),(U_2,V_2)) + J_{FMn}((x,y),(U_2,V_2))^\tau].$$

Let Σ_{FMn} be the covariance matrix of $vec(J_{FMn}^{(1)}(X,Y))$. Then

$$\sqrt{n}(vec(\hat{M}_{FMn}) - vec(M_{FMn})) \xrightarrow{\mathcal{L}} N(0,\Sigma_{FMn}), \quad \text{as } n \to \infty.$$

5 Implementation

In this section, we describe the procedures for estimating $\mathcal{S}_{Y|X}$ and $\mathcal{S}_{E[Y|X]}$ using the estimated candidate matrices $\hat{M}_{\rm FC}$ and $\hat{M}_{\rm FM}$, respectively. The determination of dimensionality q and the choice of tuning parameters $\sigma_{\rm W}^2$ and $\sigma_{\rm T}^2$ are also discussed.

5.1 Algorithms

If the dimensionality of $\mathcal{S}_{Y|X}$ (or $\mathcal{S}_{E[Y|X]}$) is known to be q, then the first q eigenvectors of $M_{\rm FC}$ (or $M_{\rm FM}$) form an orthogonal basis of $\mathcal{S}_{Y|X}$ (or $\mathcal{S}_{E[Y|X]}$). From the previous sections, it is known that the eigenvectors of $\hat{M}_{\rm FC}$ (or $\hat{M}_{\rm FM}$) converge to those of $M_{\rm FC}$ (or $M_{\rm FM}$) at the rate of $\sqrt{n}$. Therefore, we can use the first q eigenvectors of $\hat{M}_{\rm FC}$ (or $\hat{M}_{\rm FM}$) to generate a linear subspace and use it as an estimate of $\mathcal{S}_{Y|X}$ (or $\mathcal{S}_{E[Y|X]}$), which is denoted by $\hat{\mathcal{S}}_{Y|X}$ (or $\hat{\mathcal{S}}_{E[Y|X]}$). In practice, X does not necessarily have zero mean and identity covariance matrix, so the data needs to be standardized first. Standardizations generally do not affect the convergence rate of the estimates, but may change their asymptotic variances. The procedure to derive $\hat{\mathcal{S}}_{Y|X}$ (or $\hat{\mathcal{S}}_{E[Y|X]}$) is summarized as follows.

0. Specify parameters q, $\sigma_{\rm W}^2$, and $\sigma_{\rm T}^2$ ($\sigma_{\rm T}^2$ is not needed when estimating $\mathcal{S}_{E[Y|X]}$).
1. Standardize data as follows: $\tilde{x}_i = \hat{\Sigma}^{-1/2}(x_i - \bar{x})$ and $\tilde{y}_i = (y_i - \bar{y})/s_y$, where $\bar{x}$ and $\hat{\Sigma}$ are the sample mean and the sample covariance matrix of x_i's, and $\bar{y}$ and s_y are the sample mean and the standard deviation of y_i's.
2. Calculate $\hat{M}_{\rm FCn}$ (or $\hat{M}_{\rm FMn}$) using standardized data $\{(\tilde{x}_i, \tilde{y}_i)\}_{1\le i\le n}$.
3. Obtain the spectral decomposition of $\hat{M}_{\rm FCn}$ (or $\hat{M}_{\rm FMn}$), that is, the eigenvector-eigenvalue pairs $(\hat{e}_1, \hat{\lambda}_1), \ldots, (\hat{e}_p, \hat{\lambda}_p)$ with $\hat{\lambda}_1 \ge \cdots \ge \hat{\lambda}_p$.
4. Then $\hat{\mathcal{S}}_{Y|X}$(or $\hat{\mathcal{S}}_{E[Y|X]}$) $= \text{span}\{\hat{\Sigma}^{-1/2}\hat{e}_1, \ldots, \hat{\Sigma}^{-1/2}\hat{e}_q\}$.

The procedure listed above is fairly standard in sufficient dimension reduction except that the estimated candidate matrices $\hat{M}_{\rm FCn}$ (or $\hat{M}_{\rm FMn}$) based on the Fourier method is used in Step 2. For convenience, in the rest of the paper, we will simply refer to the procedure with $\hat{M}_{\rm FCn}$ as FC standing for the Fourier method for estimating Central subspace, and the procedure with $\hat{M}_{\rm FMn}$ as FM standing for the Fourier method for estimating the central Mean subspace. There are two parameters q and $\sigma_{\rm W}^2$ that need to be specified in FM, while in FC, an additional parameter $\sigma_{\rm T}^2$ needs to be chosen. Note that q is different from the other two parameters in that the former is a model parameter and the latter are tuning parameters. The procedures described above assume that these parameters are known. Next we discuss the determination of dimensionality q and the choice of the tuning parameters.

5.2 Determination of Dimensionality q

In practice, the dimension of $\mathcal{S}_{Y|X}$ (or $\mathcal{S}_{E[Y|X]}$) is unknown, and needs to be inferred from data. One informal method is to generate the scree plot of the eigenvalues of $\hat{M}_{\text{FCn}}$ (or $\hat{M}_{\text{FMn}}$) as in principal component analysis and look for an "elbow" pattern in the plot. The dimension q is chosen to be the number of dominant eigenvalues. Several more formal methods for choosing q were proposed in the literature. For example, Li (1991, 1992) proposed to use a χ^2 statistic to sequentially test $q = 0, 1, 2$, etc., while Cook and Yin (2001) advocated using permutation test for the same purpose. Recently, Ye and Weiss (2003) proposed to use the bootstrap procedure to determine q. Next we follow the basic idea of Ye and Weiss (2003) to develop the bootstrap procedure to choose q in FC (or FM).

First we introduce a distance measure for two subspaces of $\mathbb{R}^p$ and then use it to define the variability of an estimated subspace. Let A and B be two $p \times q$ matrices of full column rank and $\mathcal{S}(A)$ and $\mathcal{S}(B)$ be the column spaces of A and B respectively. Let $P_A = A(A^\tau A)^- A^\tau$ and $P_B = B(B^\tau B)^- B^\tau$ be the projection matrices onto $\mathcal{S}(A)$ and $\mathcal{S}(B)$ respectively, where $^-$ represents the generalized inverse of a matrix. We define the *trace correlation* r between $\mathcal{S}(A)$ and $\mathcal{S}(B)$ to be $r = \sqrt{\frac{1}{q} tr(P_A P_B)}$. It can be verified that $0 \le r \le 1$, and r is equal to 0 if $\mathcal{S}(A)$ and $\mathcal{S}(B)$ are perpendicular to each other; r is equal to 1 if $\mathcal{S}(A)$ and $\mathcal{S}(B)$ are identical. The larger r is, the closer $\mathcal{S}(A)$ is to $\mathcal{S}(B)$. Hence, we use $d = 1 - r$ as a metric of the distance between $S(A)$ and $S(B)$. See Ye and Weiss (2003) for more discussion.

Given $\{(x_i, y_i)\}_{1 \le i \le n}$, let $\hat{\mathcal{S}}_q$ be the estimate of $\mathcal{S}$ for a fixed q, where $\mathcal{S}$ represents $\mathcal{S}_{Y|X}$ or $\mathcal{S}_{E[Y|X]}$. The variability of $\hat{\mathcal{S}}_q$ can be evaluated by the bootstrap procedure described below.

(1) Randomly re-sample from $\{(x_i, y_i)\}_{1 \le i \le n}$ with replacement to generate N bootstrap samples each of size n, and the jth sample is denoted by $\{(x_i^{(j)}, y_i^{(j)})\}_{1 \le i \le n}$ for $1 \le j \le N$;

(2) Based on each bootstrap sample, e.g., the jth sample $\{(x_i^{(j)}, y_i^{(j)})\}_{1 \le i \le n}$, derive the estimate of $\mathcal{S}$ and denote it by $\hat{\mathcal{S}}_q^{(j)}$;

(3) Calculate the distance between $\hat{\mathcal{S}}_q^{(j)}$ and $\hat{\mathcal{S}}_q$ and denote it by $d_q^{(j)}$;

(4) Calculate $\bar{d}(q) = \frac{1}{N} \sum_{j=1}^{N} d_q^{(j)}$, which is the average distance between $\hat{\mathcal{S}}_q^{(j)}$ and $\hat{\mathcal{S}}_q$ for $1 \le j \le N$. We use $\bar{d}(q)$ as a measure of the variability of $\hat{\mathcal{S}}_q$.

Repeating the procedure above for $q = 1, \ldots, p$ results in $\{\bar{d}(q)\}_{q=1}^{p}$, which are the variabilities of $\hat{\mathcal{S}}_q$ for $1 \le q \le p$.

Suppose the true dimensionality of $\mathcal{S}$ is equal to q_0. When $q < q_0$, $\hat{\mathcal{S}}_q$ estimates a q-dimensional proper subspace of $\mathcal{S}$. Because there are infinitely many such subspaces, $\hat{\mathcal{S}}_q$ is expected to demonstrate

large variability, and the smaller q is, the larger the variability (or $\bar{d}(q)$) is. When q is slightly bigger than q_0, $\hat{\mathcal{S}}_q$ estimates $\mathcal{S} \oplus \tilde{\mathcal{S}}$, where $\tilde{\mathcal{S}}$ is a $(q - q_0)$-dimensional space orthogonal to $\mathcal{S}$. Because $\tilde{\mathcal{S}}$ can be arbitrary, $\hat{\mathcal{S}}_q$ is also expected to show large variability, that is, large $\bar{d}(q)$. When q is getting bigger and closer to p, $\hat{\mathcal{S}}_q$ estimates almost the whole space $\mathbb{R}^p$, so the variability of $\hat{\mathcal{S}}_q$ starts to decrease and eventually becomes zero. When $q = q_0$, $\hat{\mathcal{S}}_q$ and $\hat{\mathcal{S}}_q^{(j)}$ estimate the same fixed space $\mathcal{S}$, hence the variability of $\hat{\mathcal{S}}_q$ is expected to be small. In summary, $\bar{d}(q)$ demonstrates the following overall trend. It decreases for $1 \le q \le q_0$, then increases for $q_0 \le q \le q_*$, where q_* is a maximizer of $\bar{d}(q)$, then decreases to zero for $q_* \le q \le p$. We call q_0 the valley and q_* the peak of this trend. In real data analysis, it is possible to have local fluctuations that are not consistent with the overall trend. To choose q_0, we plot $\bar{d}(q)$ against q and look for the overall trend in the plot ignoring possible local deviations, then the valley is chosen to be q_0. The plot of $\bar{d}(q)$ versus q is called the *dimension variability plot*. Examples on how to use the dimension variability plot to choose q_0 will be given in Section 6.

5.3 Choice of σ_{W}^2 and σ_{T}^2

The tuning parameters σ_{W} and σ_{T} may be considered as the bandwidths of the weight functions $K(\omega)$ and $k(t)$. The selection of σ_{W} and σ_{T}, however, is fundamentally different from the selection of bandwidth for kernels used in nonparametric function estimation. In the latter case, the bandwidth of a kernel needs to decrease to zero to ensure the asymptotic consistency of the estimated function as sample size goes to infinity. In FC and FM, the consistency of $\hat{M}_{\mathrm{FM}}$ and $\hat{M}_{\mathrm{FC}}$ hold for any fixed positive σ_{W}^2 and σ_{T}^2; see Proposition 3 and Proposition 6. Nevertheless, given a finite sample, the choice of σ_{W}^2 and σ_{T}^2 affects the variability of the resulted estimates, so they need to be chosen carefully. In the following, we first discuss the heuristics for choosing σ_{W}^2 and σ_{T}^2 and give their recommended values, then we briefly introduce a bootstrap procedure for their optimal selection again following the idea of Ye and Weiss (2003).

Let us discuss σ_{W}^2 first. When σ_{W}^2 is too large, $\phi(\omega, t)$ with large $\|\omega\|$ will receive much larger weight than when σ_{W}^2 is small. As explained earlier, $\phi(\omega, t)$ with large $\|\omega\|$ corresponds to patterns with high frequencies, which may not be as important as the patterns with low frequencies and are sensitive to noise. Therefore, large σ_{W}^2 makes FC unstable, especially when the sample size is moderate. On the other hand, if σ_{W}^2 is too small, for example, close to zero, then the weight assigned to $\phi(\omega, t)$ is almost zero except for ω in a small neighborhood of the origin. By Proposition 8, FC is close to SIR when σ_{W}^2 is small, and may miss some symmetric directions. Hence, we need to use a value of σ_{W}^2 that is neither too large nor too small. Based on our empirical study, we have found that $\sigma_{\mathrm{W}} = 1/3$, or equivalently, $\sigma_{\mathrm{W}}^2 = 0.1$, generally works well for standardized data. Similarly for FM, we also recommend to use

$\sigma_{\rm W}^2 = 0.1$ in calculating $\hat{M}_{\rm FM}$.

The interpretation of $\sigma_{\rm T}^2$ is slightly different from that of $\sigma_{\rm W}^2$. Theoretically, we use $k(t)$ to pool the central mean subspaces $\mathcal{S}_{E[T(Y,t)|X]}$ together to recover the entire $\mathcal{S}_{Y|X}$. When $\sigma_{\rm T}^2 = 0$, $\mathcal{S}(M_{\rm FC})$ degenerates to the null space. When $\sigma_{\rm T}^2$ is too large, a relatively large amount of weight is assigned to the central mean subspaces $\mathcal{S}_{E[T(Y,t)|X]}$ with large t, which corresponds to features of the response with high frequencies. This would make FM unstable and sensitive to noise. Hence, we need to use a value of $\sigma_{\rm T}^2$ that is neither too large nor too small. Our extensive empirical study suggests that $\sigma_{\rm T}^2 = 1$ be a good choice for standardized response.

The recommendations for $\sigma_{\rm W}^2$ and $\sigma_{\rm T}^2$ above are based on heuristics and empirical study. A more formal approach is to use the bootstrap procedure introduced in the previous subsection. Let us use the choice of $\sigma_{\rm W}^2$ as an illustration. First we choose m candidate values $\sigma_1^2, \ldots, \sigma_m^2$, which are usually equally spaced in a given interval. For any σ_i^2, $i = 1, \ldots, m$, calculate $\bar{d}(\sigma_i^2)$ using a bootstrap procedure similar to that described in the previous subsection. Note that the dimensionality q is assumed to be fixed, instead, $\sigma_{\rm W}^2$ is varied here. The optimal $\sigma_{\rm W}^2$ is chosen to be the σ_i^2 that minimizes $\bar{d}(\sigma_i^2)$. The optimal $\sigma_{\rm T}^2$ can be obtained using a similar bootstrap procedure.

In practice, some applications may require to optimally choose q, $\sigma_{\rm W}^2$ and $\sigma_{\rm T}^2$. We recommend the following procedure to use bootstrap repeatedly. First, q is determined with $\sigma_{\rm W}^2 = 0.1$ and $\sigma_{\rm T}^2 = 1.0$ and is denoted by q_1; second, $\sigma_{\rm W}^2$ is chosen with $\sigma_{\rm T}^2 = 1.0$ and q_1 and is denoted by $\sigma_{\rm W1}^2$; third, $\sigma_{\rm T}^2$ is selected with $\sigma_{\rm W1}^2$ and q_1 and denoted by $\sigma_{\rm T1}^2$. The steps are iterated until the parameters are stabilized. Our experiences suggest that one iteration is usually sufficient. The second and third steps can in fact be combined if a two-dimensional grid is employed for selecting $\sigma_{\rm W}^2$ and $\sigma_{\rm T}^2$ simultaneously.

6 Examples

In this section, we present four examples to demonstrate the performance of FC and FM and compare them with other existing methods. The first three examples are based on synthetic models where $X = (X_1, \ldots, X_{10})^\tau$ denotes a random vector in $\mathbb{R}^{10}$, ε denotes a random error, X_i's and ε are independent and identically distributed as $N(0,1)$, $\beta_1 = (1,1,1,1,0,0,0,0,0,0)^\tau$ and $\beta_2 = (0,0,0,0,0,0,1,1,1,1)^\tau$. In the last example, we apply FC to a real data set called 1985 Automobile Data to study how the price of car depends on its features.

Example 1. Consider the model $Y = (\beta_1^\tau X)^2/(3 + (\beta_2^\tau X + 2)^2) + 0.2\varepsilon$. In this model the central subspace and the central mean subspace are identical, that is, $\mathcal{S}_{Y|X} = \mathcal{S}_{E[Y|X]} = \mathcal{S}(\beta_1, \beta_2)$. Let $\mathcal{S} = \mathcal{S}(\beta_1, \beta_2)$. So, no matter whether a method is targeting the central mean subspace or the central

subspace, it can be used to estimate $\mathcal{S}$. The dimension of $\mathcal{S}$, $q = 2$, is assumed to be known. We compare FC ($\sigma_{\mathrm{W}}^2 = 0.1$, $\sigma_{\mathrm{T}}^2 = 1.0$) and FM ($\sigma_{\mathrm{W}}^2 = 0.1$) with other five existing methods including SIR (five slices), SAVE (five slices), y-pHd, r-pHd and IHT as follows. We randomly generate 500 samples of size $n = 500$ from the model. For each sample, we apply the seven methods listed above one by one to obtain the estimates of $\mathcal{S}$. Then we calculate the distances between these estimates and $\mathcal{S}$. For each method, we generate a boxplot for the 500 distances. This procedure results in seven boxplots that are displayed side-by-side in Figure 1. From Figure 1, we conclude that both FC and

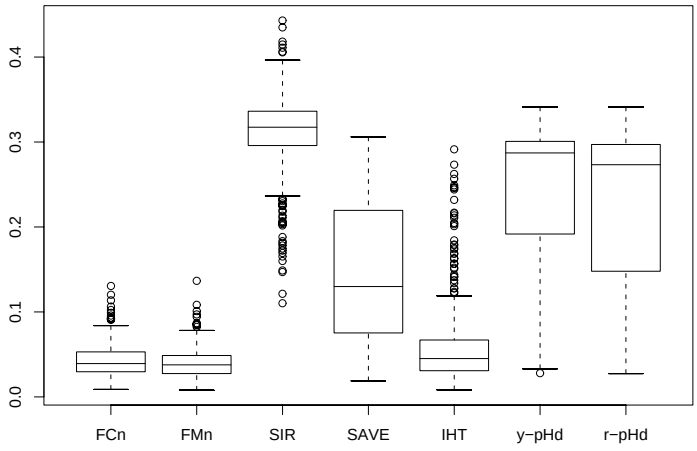

Figure 1: Side-by-side boxplots for the performance comparison between FC, FM, SIR, SAVE, IHT, y-pHd and r-pHd in Example 1. The y-axis represents the distance between an estimated subspace and the true subspace. Each boxplot is based on 500 samples.

FM outperform the other methods. IHT has similar performance as FC and FM, but demonstrates slightly larger variability. A possible explanation for IHT's good performance is that it is carefully designed to capture the monotone and U-shaped trends in the function that links $E[Y|X]$ and X.

Example 2. Consider the following heteroscedastic model, $Y = (\beta_1^\tau X) + 4\,(\beta_2^\tau X)\,\varepsilon$. For this model, the central mean subspace and the central subspace are different, because $\mathcal{S}_{E[Y|X]} = \mathcal{S}(\beta_1)$ and $\mathcal{S}_{Y|X} = \mathcal{S}(\beta_1, \beta_2)$. Clearly $\mathcal{S}_{E[Y|X]}$ is only a proper subspace of $\mathcal{S}_{Y|X}$, so we focus our attention on $\mathcal{S}_{Y|X}$ and the methods aimed at estimating $\mathcal{S}_{Y|X}$. We draw a sample of 500 data points, and apply FC ($\sigma_{\mathrm{W}}^2 = 0.1$ and $\sigma_{\mathrm{T}}^2 = 1.0$) to estimate $\mathcal{S}_{Y|X}$. Only the first two eigenvalues of the estimated candidate matrix $\hat{M}_{\mathrm{FCn}}$ are relatively large. We use the bootstrap procedure to generate the dimension variability plot based on 500 bootstrap samples, which is included as the left plot in Figure 2. Using the rule described in Section 5.2, it confirms that the dimension of $\mathcal{S}_{Y|X}$ is equal to 2. Therefore, $\mathcal{S}_{Y|X}$ is estimated by

the space spanned the first two eigenvectors of $\hat{M}_{\mathrm{FCn}}$, which are

$$\hat{\beta}_1^\tau = (0.5012, 0.4414, 0.4869, 0.4015, -0.0205, 0.2531, -0.1111, 0.0366, -0.0709, 0.0266),$$
$$\hat{\beta}_2^\tau = (0.0381, 0.2085, -0.0590, 0.0025, -0.1442, -0.0137, 0.4853, 0.5143, 0.5186, 0.3658).$$

The distance between $\hat{\mathcal{S}}_{Y|X} = \mathcal{S}(\hat{\beta}_1, \hat{\beta}_2)$ and the true $\mathcal{S}_{Y|X}$ is 0.04388.

In order to compare FC with SIR (five slices) and SAVE (five slices), we draw 500 samples of size $n = 500$ from the model, apply the methods to the samples to obtain the estimated central subspaces, and calculate the distances between the estimated central subspaces and $\mathcal{S}_{Y|X}$. The distances are summarized by the side-by-side boxplots included in the right plot of Figure 2. The boxplots indicates that FC outperforms both SIR and SAVE in this example.

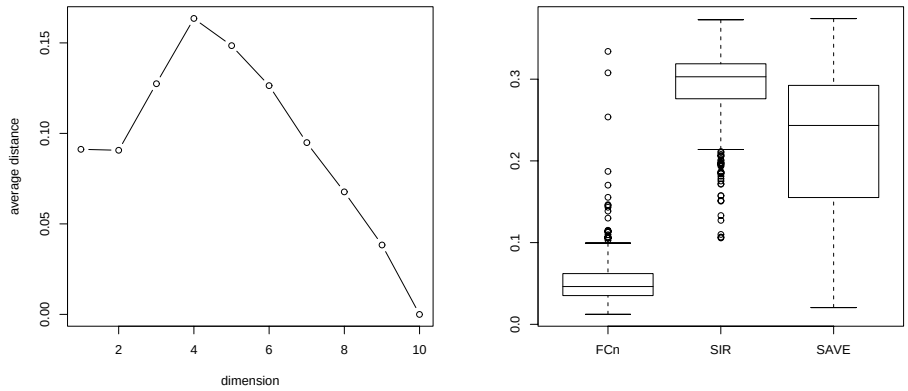

Figure 2: On the left is the dimension variability plot of $\bar{d}(q)$ versus q based on 500 bootstrap samples in Example 2. On the right is the side-by-side boxplots for the performance comparison between FC, SIR and SAVE in Example 2. The y-axis represents the distance between an estimated subspace and the true subspace. Each boxplot is based on 500 samples.

Example 3. Consider the following model with discrete response, $Y = I_{[\beta_1^\tau X + \sigma\varepsilon > 1]} + 2I_{[\beta_2^\tau X + \sigma\varepsilon > 0]}$, where $I_{[\cdot]}$ denotes the indicator function and $\sigma = 0.2$. So the possible values for Y are 0, 1, 2, and 3. In this example we only consider the central subspace, which is spanned by β_1 and β_2. It is clear that Y is not a continuous function of X. In the derivation of FC, a few differentiability conditions are required. But in the final formula for M_{FC}, no differentiation is involved. So we expect FC would still work in this example. As a matter of fact, the involved differentiability in deriving FC is required only in the generalized sense. We draw a sample of 500 points, and apply FC ($\sigma_{\mathrm{W}}^2 = 0.1$ and $\sigma_{\mathrm{T}}^2 = 3.0$) to estimate $\mathcal{S}_{Y|X}$. Note that σ_{T}^2 is chosen to be larger than the usual recommended value, because the discontinuity in the model represents a feature with high frequency and it could

not be well captured by the transformed response $\exp\{\imath t y\}$ with small t. The first two eigenvalues of $\hat{M}_{\mathrm{FCn}}$ are relatively large indicating that the dimension of $\mathcal{S}_{Y|X}$ is 2. This is further confirmed by the dimension variability plot included in Figure 3. Therefore, the estimated central subspace is spanned by the first two eigenvectors of $\hat{M}_{\mathrm{FCn}}$, which are,

$$\hat{\beta}_1^\tau = (0.0603, 0.0927, 0.0567, 0.0521, -0.0432, -0.0106, 0.4964, 0.3776, 0.4923, 0.4571),$$
$$\hat{\beta}_2^\tau = (0.5274, 0.5209, 0.5306, 0.4528, 0.0580, 0.1041, -0.0735, -0.0771, -0.0729, -0.0819).$$

The distance between $\mathcal{S}(\hat{\beta}_1, \hat{\beta}_2)$ and the true subspace $\mathcal{S}_{Y|X}$ is $d = 0.007809$.

We use the same procedure as in Example 2 to compare FC ($\sigma_{\mathrm{W}}^2 = 0.1$ and $\sigma_{\mathrm{T}}^2 = 3.0$) with SIR (four slices) and SAVE (four slices). The three boxplots corresponding to FC, SIR and SAVE are included as the right plot in Figure 3. In this example, the performances of the three methods are comparable with SIR slightly better than the other two. One explanation for the better performance of SIR is that SIR can be regarded as an extension of linear discriminant analysis.

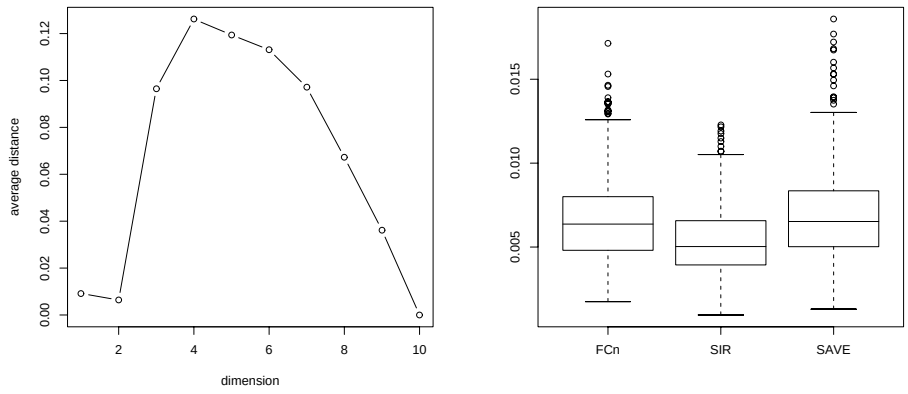

Figure 3: On the left is the dimension variability plot based on 500 bootstrap samples in Example 3. On the right is the side-by-side boxplots for the performance comparison between FC, SIR and SAVE in Example 3. The y-axis represents the distance between an estimated subspace and the true subspace. Each boxplot is based on 500 samples.

Example 4. In this example, we use FC to analyze a real data set titled 1985 Automobile Data. The objective is to study how the price of car depends on its features. The data set is available at the UCI Machine Learning Repository (`ftp://ftp.ics.uci.edu/pub/machine-learning-databases/autos`). Originally, there are 205 instances (or cases), 26 attributes (or variables) in the data set and there are also some missing values. Because most current dimension reduction methods including FC can only handle continuous variables, we remove eight categorical variables from the data set. We remove one

continuous variable that contains many missing values. For simplicity, we also discard the instances with missing values. The resulted data contains 195 instances and 14 variables, which are Wheelbase (x_1), Length (x_2), Width (x_3), Height (x_4), Curb weight (x_5), Engine size (x_6), Bore (x_7), Stroke (x_8), Compression ratio (x_9), Horsepower (x_{10}), Peak rpm (x_{11}), City mpg (x_{12}), Highway mpg (x_{13}) and Price (y). We use the logarithm of Price ($\log(y)$) as the response and x_1 to x_{13} as the predictors. Before we apply FC to the data, we standardize each predictors using their means and standard deviations.

We first use FC with $\sigma_{\mathrm{W}}^2 = 0.1$ and $\sigma_{\mathrm{T}}^2 = 1.0$ and the bootstrap procedure to choose the dimension of the central subspace. The results show that the dimension should be two. In order to obtain sharper views, we fix $q = 2$, and further use the bootstrap procedure mentioned in the end of Section 5 to tune the parameters σ_{W}^2 and σ_{T}^2. We have found that $\sigma_{\mathrm{W}}^2 = .08$ and $\sigma_{\mathrm{T}}^2 = .6$ are better choices. Employing FC with the tuned parameters, we calculate $\hat{M}_{\mathrm{FCn}}$ and derive its spectral decomposition. The first two eigenvalues of $\hat{M}_{\mathrm{FCn}}$ are dominant, which are $\hat{\lambda}_1 = 0.1071$ and $\hat{\lambda}_2 = 0.0381$. The bootstrap procedure is run with 1000 bootstrap samples and the dimension variability plot is generated and included as the left plot in Figure 5. The plot suggests that $q = 2$. So, we use the first two eigenvectors of $\hat{M}_{\mathrm{FCn}}$ to derive the estimate $\hat{\mathcal{S}}_{Y|X} = \mathcal{S}(\hat{\beta}_1, \hat{\beta}_2)$, where $\hat{\beta}_1^\tau = (0.05, -0.19, 0.08, 0.09, 0.75, -0.24, 0.00, -0.11, 0.17, 0.51, 0.04, -0.08, 0.12)$ and $\hat{\beta}_2^\tau = (0.08, -0.38, 0.09, 0.08, 0.03, 0.70, -0.17, -0.20, -0.06, -0.08, 0.06, 0.49, -0.17)$. Figure 4 includes the projection plots of $\log(y)$ versus $\hat{\beta}_1^\tau x$ (left) and $\log(y)$ versus $\hat{\beta}_2^\tau x$ (right), with the left displaying a strong linear relationship and the right displaying a parabolic relationship.

Based on the relative magnitudes of the components, $\hat{\beta}_1$ is mainly determined by Curb weight (x_5) and Horsepower (x_{10}) followed by the other variables, and $\hat{\beta}_2$ is mainly determined by Engine size (x_6) and City mpg (x_{12}). The strong linear relationship between $\log(y)$ and $\hat{\beta}_1^\tau x$ indicates that the price of a car can be well predicted by its Curb weight and Horsepower. The parabolic relationship between $\log(y)$ and $\hat{\beta}_2^\tau x$ reveals that the price of a car also depends on its Engine size and City mpg, however, in a slightly more complicated manner. After checking the makes and styles of the cars reported in the original data, we have found that the points in the upper branch of the parabola represent high-end cars such as the sedans or the convertibles of Mercedes-Benz, BMW, Jaguar, Porche etc., while the points in the lower branch represent lower-end cars such as the hatchbacks of Honda, Chervolet, Polymouth, Subaru, etc. For the high-end cars, the price increases as $\hat{\beta}_2^\tau x$ increases, while for the lower-end cars, the price decreases as $\hat{\beta}_2^\tau x$ increases. The two relationships above further imply that there exists a nonlinear confounding between $\hat{\beta}_1^\tau x$ and $\hat{\beta}_2^\tau x$. The right plot in Figure 5 is the plot of $\hat{\beta}_1^\tau x$ versus $\hat{\beta}_2^\tau x$, which exhibits a parabolic relationship between these two directions. Readers can consult

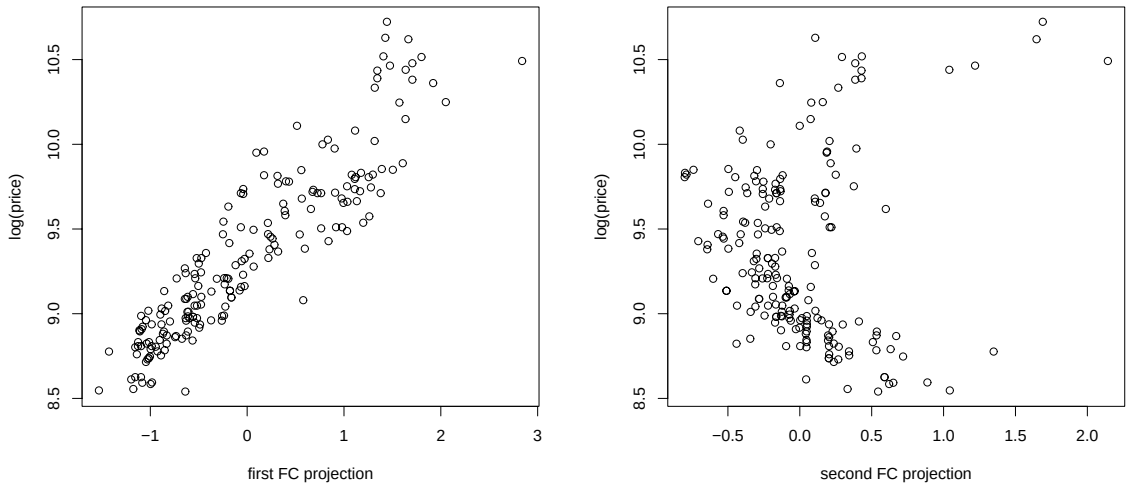

Figure 4: On the left is the plot of $\log(y)$ versus $\hat{\beta}_1^\tau x$ and on the right is the plot of $\log(y)$ versus $\hat{\beta}_2^\tau x$ in Example 4.

Li (1997) for a detailed discussion of nonlinear confounding between predictors. SIR and SAVE are also used to analyze the data. The former gives similar results as FC, while the latter fails to recover interesting directions in this particular example.

7 Conclusion

Using the Fourier transform, we have derived two candidate matrices $M_{\rm FC}$ and $M_{\rm FM}$ to recover the entire central subspace and central mean subspace respectively. Under further distributional assumptions, explicit estimates of the candidate matrices are derived, which lead to the estimates of the central and central mean subspaces. The selection of the tuning parameters and the determination of dimensionality have been discussed. Synthetic and real examples are used to demonstrate the performances of the proposed methods in comparison with other existing ones. The use of the Fourier transform may provide a different view on dimension reduction in general regression, which is expected to generate more interesting results in the future. Currently, we are focused on two issues. The first one is to generalize the results of the current paper to the case without distributional assumptions imposed on X, and the second is to develop dimension reduction techniques for regression with multiple responses. Because our approach does not involve the slicing of the response and the partition of data, it may be more appropriate for the latter case than SIR and SAVE.

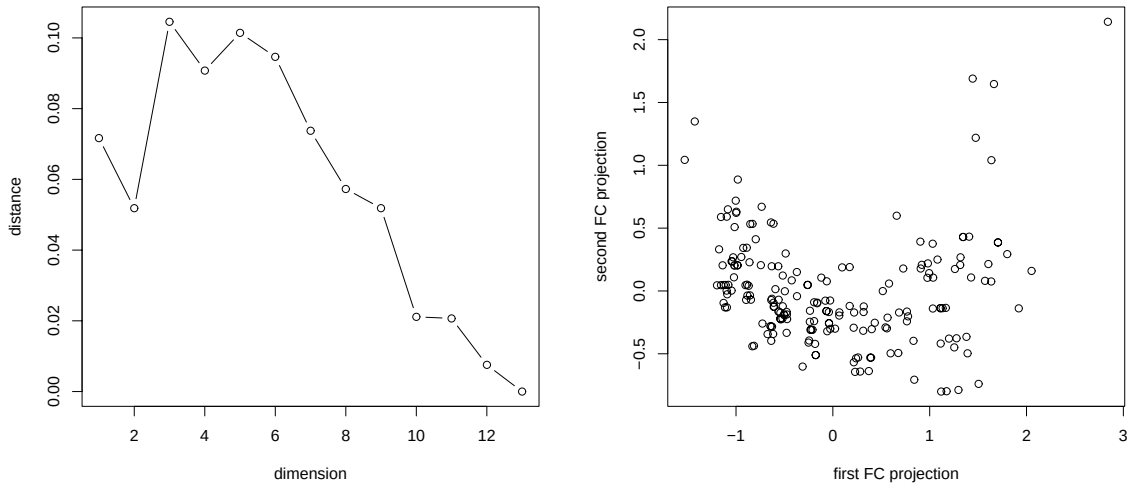

Figure 5: On the left is the dimension variability plot generated from the bootstrap procedure with 1000 bootstrap samples, $\sigma_{\mathrm{W}}^2 = 0.08$ and $\sigma_{\mathrm{T}}^2 = 0.6$; on the right is the plot of $\hat{\beta}_1^\tau x$ versus $\hat{\beta}_2^\tau x$ that shows a nonlinear relationship. Both plots are for Example 4.

8 Appendix

Proof of Equation (5). Because $A = (\alpha_1, \alpha_2, \ldots, \alpha_q)$ and the columns $\alpha_1, \ldots, \alpha_q$ form a basis for $\mathcal{S}_{E[Y|X]}$ with q dimensions, to prove equation (5), it is sufficient to show that for $\beta \in \mathbb{R}^p$, $\beta^\tau A = 0$ is equivalent to $\beta^\tau \frac{\partial}{\partial x} m(x) = 0$ for all $x \in \mathrm{supp}(X)$.

By the chain rule of differentiation, $\frac{\partial}{\partial x} m(x) = A \frac{\partial}{\partial u} g(u)$. So $\beta^\tau A = 0$ immediately implies that $\beta^\tau \frac{\partial}{\partial x} m(x) = \beta^\tau A \frac{\partial}{\partial u} g(u) = 0$ for all $x \in \mathrm{supp}(X)$.

Next, we show the converse is also true by contradiction. Assume that there exists $\beta_0 \in \mathbb{R}^p$ such that $\beta_0^\tau \frac{\partial}{\partial x} m(x) = 0$ for all $x \in \mathrm{supp}(X)$, but $\beta_0^\tau A \neq 0$. Let $\xi_1 = A^\tau \beta_0 / \|A^\tau \beta_0\|$. Clearly, ξ_1 is a q-dimensional nonzero vector. So $\beta_0^\tau \frac{\partial}{\partial x} m(x) = \beta_0^\tau A \frac{\partial}{\partial u} g(u) = 0$ means $\xi_1^\tau \frac{\partial}{\partial u} g(u) = 0$, which implies that the directional derivative of g as a function of $u = (u_1, \ldots, u_q)^\tau$ along ξ_1 is always zero. This further implies that $g(u)$ is a constant along ξ_1, that is, $g(u + t\xi_1) = g(u)$ for $t \in \mathbb{R}$. We can expand ξ_1 by bringing in $\xi_2, \ldots, \xi_q$ to form an orthonormal basis for $\mathbb{R}^q$. Let $D = (\xi_1, \ldots, \xi_q)$ and $v = D^\tau u = (v_1, \ldots, v_q)^\tau$, then $g(u) = g(Dv)$ and $\frac{\partial}{\partial v_1} g(Dv) = \xi_1^\tau \frac{\partial}{\partial u} g(u) = 0$. Hence g does not depend on v_1, so we can rewrite $g(u) = g(Dv) = \tilde{g}(v_2, \ldots, v_q) = \tilde{g}(\xi_2^\tau A^\tau x, \ldots, \xi_q^\tau A^\tau x)$, which implies that $\mathcal{S}(A\xi_2, \ldots, A\xi_q)$ is also a dimension reduction subspace for $E[Y|X]$, and the central mean subspace has dimension at most $q - 1$, which contradicts that $\dim(\mathcal{S}_{E[Y|X]}) = q$. Thus the proof is completed. □

Proof of Proposition 1.

1. From the definition of $\psi(\omega)$ in (6), $a(\omega)$ and $b(\omega)$ have the following expression

$$a(\omega) = \int \cos(\omega^\tau x)\big(\tfrac{\partial}{\partial x}\, m(x)\big)\, f_X(x)\, dx$$
$$b(\omega) = \int \sin(\omega^\tau x)\big(\tfrac{\partial}{\partial x}\, m(x)\big)\, f_X(x)\, dx$$

   From (5), $\frac{\partial}{\partial x}\, m(x) \in \mathcal{S}_{E[Y|X]}$, for all $x \in \text{supp}(X)$. So both $a(\omega)$ and $b(\omega)$ belong to $\mathcal{S}_{E[Y|X]}$. Furthermore, by (7), the inverse Fourier transform equation, and the fact that the Fourier transform is one-to-one, we have $\beta^\tau(\frac{\partial}{\partial x}m(x))f_X(x) = 0$ for all $x \in \mathbb{R}^p$ is equivalent to $\beta^\tau\psi(\omega) = 0$, for all $\omega \in \mathbb{R}^p$, which is further equivalent to $\beta^\tau a(\omega) = \beta^\tau b(\omega) = 0$ for all $\omega \in \mathbb{R}^p$. Using (5), we have $\mathcal{S}_{E[Y|X]} = \text{span}\{a(\omega), b(\omega) : \omega \in \mathbb{R}^p\}$.

2. Because $m(x)f_X(x) \to 0$ as $\|x\| \to \infty$. Using integration by parts, we have

$$\begin{aligned}\psi(\omega) &= -\int m(x)\,(\imath\omega \exp\{\imath\omega^\tau x\} f_X(x) + \exp\{\imath\omega^\tau x\}\tfrac{\partial}{\partial x} f_X(x))\, dx \\ &= -E_{(X,Y)}[Y(\imath\omega + G(X))\, \exp\{\imath\omega^\tau X\}].\end{aligned}$$

3. The proof is standard and can be found in Folland (1992; Pages 217, 243).

4. The proof is standard and can be found in Folland (1992; Pages 222, 244).

□

Proof of Proposition 2. Because $\psi(\omega) = a(\omega) + \imath b(\omega)$, we have

$$\psi(\omega)\bar{\psi}(\omega)^\tau = [a(\omega)a(\omega)^\tau + b(\omega)b(\omega)^\tau] + \imath[b(\omega)a(\omega)^\tau - a(\omega)b(\omega)^\tau].$$

By (9), we know that $\int[b(\omega)a(\omega)^\tau - a(\omega)b(\omega)^\tau]d\omega = 0$. Therefore,

$$M^*_{\text{FM}} = (2\pi)^{-p}\int \psi(\omega)\bar{\psi}(\omega)^\tau\, d\omega = (2\pi)^{-p}\int [a(\omega)a(\omega)^\tau + b(\omega)b(\omega)^\tau]\, d\omega.$$

Clearly M^*_{FM} is a real non-negative definite matrix. For any p-dimensional vector β, $\beta^\tau M^*_{\text{FM}}\beta = 0$ is equivalent to $\beta^\tau a(\omega) = \beta^\tau b(\omega) = 0$ for all ω, which implies that the column space of M^*_{FM}, $\mathcal{S}(M^*_{\text{FM}})$, is the same as $\text{span}\{a(\omega), b(\omega) : \omega \in \mathbb{R}^p\}$. By the first property in Proposition 1, we have $\mathcal{S}(M^*_{\text{FM}}) = \mathcal{S}_{E[Y|X]}$. □

Proof of Proposition 3. Because

$$M_{\text{FM}} = \int [a(\omega)a(\omega)^\tau + b(\omega)b(\omega)^\tau]\, K(\omega)\, d\omega,$$

so M_{FM} is a non-negative definite matrix. Because $K(\omega) > 0$ for $\omega \in \mathbb{R}^p$, for $\beta \in \mathbb{R}^p$, we have

$$\beta^\tau M_{\text{FM}}\beta = 0 \quad \Longleftrightarrow \quad \beta^\tau a(\omega) = \beta^\tau b(\omega) = 0 \text{ for all } \omega \in \mathbb{R}^p,$$

where $\Longleftrightarrow$ reads "being equivalent to". Therefore $\mathcal{S}(M_{\text{FM}}) = \text{span}\{a(\omega), b(\omega) : \omega \in \mathbb{R}^p\} = \mathcal{S}_{E[Y|X]}$. □

Proof of Proposition 4. Under the model assumption, following the similar argument as in the proof of (5), the central subspace can be written as

$$\mathcal{S}_{Y|X} = \text{span}\{\tfrac{\partial}{\partial x} f_{Y|X}(y \mid x) : (x, y) \in \text{supp}(X, Y)\},$$

where $f_{Y|X}$ is the conditional density of Y given X. In order to prove this proposition, it is sufficient to show that for $\beta \in \mathbb{R}^p$ and $x \in \text{supp}(X)$,

$$\beta^\tau \tfrac{\partial}{\partial x} m(x, t) = 0 \text{ for all } t \in \mathbb{R} \quad \Longleftrightarrow \quad \beta^\tau \tfrac{\partial}{\partial x} f_{Y|X}(y \mid x) = 0 \text{ for all } y \in \text{supp}(Y).$$

This is an immediate result from the fact that $\frac{\partial}{\partial x} m(x, t)$ is the Fourier transform of $\frac{\partial}{\partial x} f_{Y|X}(y \mid x)$, that is,

$$\tfrac{\partial}{\partial x} m(x, t) = \tfrac{\partial}{\partial x} \int \exp\{\imath t y\} f_{Y|X}(y \mid x)\, dy = \int \exp\{\imath t y\}\, \tfrac{\partial}{\partial x} f_{Y|X}(y \mid x)\, dy.$$

Thus the proposition is proved. □

Proof of Proposition 5. The proof is a straightforward modification of that of Proposition 1 with Y replaced by $\exp\{\imath t Y\}$ and $m(x)$ replaced by $m(x, t)$. The details are thus omitted. □

Proof of Proposition 6. Because

$$M_{\text{FC}} = \iint [a(\omega, t) a(\omega, t)^\tau + b(\omega, t) b(\omega, t)^\tau]\, K(\omega) k(t)\, d\omega dt,$$

so M_{FC} is a non-negative definite matrix. Because $K(\omega) > 0$ for $\omega \in \mathbb{R}^p$ and $k(t) > 0$ for $t \in \mathbb{R}$, we have, for any $\beta \in \mathbb{R}^p$,

$$\beta^\tau M_{\text{FC}} \beta = 0 \quad \Longleftrightarrow \quad \beta^\tau a(\omega, t) = \beta^\tau b(\omega, t) = 0 \,\text{for all}\, \omega \in \mathbb{R}^p \text{ and for all } t \in \mathbb{R}.$$

Therefore $\mathcal{S}(M_{\text{FC}}) = \text{span}\{a(\omega, t), b(\omega, t) : \omega \in \mathbb{R}^p, t \in \mathbb{R}\}$. Using the first property of Proposition 5, we have $\mathcal{S}(M_{\text{FC}}) = \mathcal{S}_{E[Y|X]}$, and the proposition is proved. □

Proof of Proposition 7. Applying integration by parts and the condition that $f_{(X,Y)}(x, y)$ goes to zero as x goes to infinity and y is fixed, we have

$$\begin{aligned}
\eta(y, \omega) &= -\int (\imath\omega + G(x))\, \exp\{\imath\omega^\tau x\}\, f_{X|Y}(x|y)\, dx \\
&= -\int \imath\omega\, \exp\{\imath\omega^\tau x\}\, f_{X|Y}(x|y)\, dx - f_Y^{-1}(y) \int \exp\{\imath\omega^\tau x\} \left(\tfrac{\partial}{\partial x} f_X(x)\right) f_{Y|X}(y|x)\, dx \\
&= f_Y^{-1}(y) \int \exp\{\imath\omega^\tau x\}\, f_X(x)\, \tfrac{\partial}{\partial x} f_{Y|X}(y|x)\, dx.
\end{aligned}$$

Since $\frac{\partial}{\partial x} f_{Y|X}(y \mid x) \in \mathcal{S}_{Y|X}$, we have $\eta(y, \omega) \in \mathcal{S}_{Y|X}$. The second part of the proposition can be easily verified. □

Proof of Proposition 8. When X follows the normal distribution with mean 0 and covariance matrix I_p, $G(X) = -X$. Because $K(\omega)$ is a point mass at $\omega = 0$,

$$M_{\text{FC}} = \int [a(0,t)a(0,t)^\tau + b(0,t)b(0,t)^\tau]k(t)\,dt,$$

where $a(0,t)$ and $b(0,t)$ are the real and imaginary parts of $\phi(0,t)$ respectively. So $\mathcal{S}(M_{\text{FC}}) = \text{span}\{a(0,t), b(0,t) : t \in \mathbb{R}\}$. By (16), $\phi(0,t)$ is the Fourier transform of $E[X|Y = y]f_Y(y)$, that is,

$$\phi(0,t) = \int \exp\{\imath ty\}\, x\, f_{(X,Y)}(x,y)\,dxdy = \int \exp\{\imath ty\}\, E[X|Y=y]\, f_Y(y)\,dy.$$

Then for any $\beta \in \mathbb{R}^p$, $\beta^\tau E[X|Y = y]f_Y(y) = 0$ for $y \in \mathbb{R}$ is equivalent to $\beta^\tau a(0,t) = \beta^\tau b(0,t) = 0$ for $t \in \mathbb{R}$. Hence $\mathcal{S}(M_{\text{FC}}) = \text{span}\{E[X|Y = y] : y \in \text{supp}(Y)\}$. The latter is exactly the space aimed at by SIR, which is $\mathcal{S}(M_{\text{SIR}})$. Thus, when $\sigma^2_{\text{W}} = 0$, $\mathcal{S}(M_{\text{FC}}) = \mathcal{S}(M_{\text{SIR}})$, and the proposition is proved. □

Proof of Theorem 1. Because $\hat{M}_{\text{FCn}}$ is a V-statistic, it can be written as follows.

$$\begin{aligned}\hat{M}_{\text{FCn}} = \frac{n-1}{2n}\binom{n}{2}^{-1} \sum_{i<j} \left\{J_{\text{FCn}}((x_i,y_i),(x_j,y_j)) + J^\tau_{\text{FCn}}((x_i,y_i),(x_j,y_j))\right\} \\ + \frac{1}{n^2}\sum_{i=1}^n J_{\text{FCn}}((x_i,y_i),(x_i,y_i)).\end{aligned}$$

The first term in the right side of the equation above is a U-statistic, and the second term is of order $O_p(n^{-1})$. Using the Hoeffding decomposition of U-statistic, we can further write $\hat{M}_{\text{FCn}}$ as

$$\begin{aligned}\hat{M}_{\text{FCn}} &= \frac{n-1}{2n}\big(2M_{\text{FCn}} + \frac{2}{n}\sum_{i=1}^n \big(J^{(1)}_{\text{FCn}}(x_i,y_i) - 2M_{\text{FCn}}\big) + o_p(n^{-1/2})\big) + O_p(n^{-1}) \\ &= M_{\text{FCn}} + \frac{1}{n}\sum_{i=1}^n \big(J^{(1)}_{\text{FCn}}(x_i,y_i) - 2M_{\text{FCn}}\big) + o_p(n^{-1/2}).\end{aligned}$$

The second term in the expression above is the average of n independent and identically distributed random matrices $(J^{(1)}_{\text{FCn}}(x_i,y_i) - 2M_{\text{FCn}})$ with $1 \le i \le n$. Because Σ_{FCn} exists, which is guaranteed by the existence of the covariance matrix of $vec(J_{\text{FCn}}((U_1,V_1),(U_2,V_2)))$, by the Central Limit Theorem, as n goes to infinity,

$$\sqrt{n}(vec(\hat{M}_{\text{FCn}}) - vec(M_{\text{FCn}})) \quad \xrightarrow{\mathcal{L}} \quad N(0, \Sigma_{\text{FCn}})$$

where $\Sigma_{\text{FCn}} = cov[vec(J^{(1)}_{\text{FCn}}(X,Y))]$ is a $p^2 \times p^2$ matrix. □

Proof of Proposition 9. Because $\phi(\omega, t) = E[\eta(Y,\omega)\exp\{\imath t Y\}]$, it is enough to prove that $\eta(Y,\omega) \in \mathcal{S}_{Y|X}$ under the weak normality condition. Applying integration by parts, we have

$$\begin{aligned}\eta(y,\omega) &= \int \exp\{\imath\omega^\tau x\}\, \tfrac{\partial}{\partial x} f_{X|Y}(x|y)\, dx + \int x\, \exp\{\imath\omega^\tau x\}\, f_{X|Y}(x|y)\, dx \\ &= f_Y(y)^{-1} \int \exp\{\imath\omega^\tau x\}\, (\tfrac{\partial}{\partial x} f_X(x) + x f_X(x))\, f_{Y|X}(y|x)\, dx \\ &\quad + f_Y(y)^{-1} \int \exp\{\imath\omega^\tau x\}\, (\tfrac{\partial}{\partial x} f_{Y|X}(y|x))\, f_X(x)\, dx.\end{aligned}$$

The second term belongs to $\mathcal{S}_{Y|X}$, so we only need to focus on the first one and denote it by I_1. Let $u_1 = B^\tau x$ and $u_2 = \tilde{B}^\tau x$, and $u = \binom{u_1}{u_2} = (B, \tilde{B})^\tau x = M^\tau x$. Then

$$\begin{aligned}I_1 &= f_Y^{-1}(y) \int \exp\{\imath\tilde{\omega}^\tau u\}\, M(\tfrac{\partial}{\partial u}\tilde{p}(u) + u\tilde{p}(u))\, p(y|u_1)\, du \\ &= f_Y^{-1}(y) B \int \exp\{\imath\tilde{\omega}^\tau u\}\, (\tfrac{\partial}{\partial u_1}\tilde{p}(u) + u_1\tilde{p}(u))\, p(y|u_1)\, du \\ &\quad + f_Y^{-1}(y)\tilde{B} \int \exp\{\imath\tilde{\omega}^\tau u\}\, (\tfrac{\partial}{\partial u_2}\tilde{p}(u) + u_2\tilde{p}(u))\, p(y|u_1)\, du_1 du_2,\end{aligned}$$

where $\tilde{\omega} = M^\tau\omega$ and $\tilde{p}(u) = f_X(Mu)$ is the density function of u. Notice that the first term falls in $\mathcal{S}_{Y|X}$. Under the weak normality condition, we have $\frac{\partial}{\partial u_2}\tilde{p}(u_2|u_1) + u_2\tilde{p}(u_2|u_1) = 0$, so the second term in the expression above is zero. Therefore $I_1 \in \mathcal{S}_{Y|X}$. This proves the proposition. □

Proof of Theorem 2. This proof is similar to that of Theorem 1 and is thus omitted. □

9 Reference

Brillinger, D. R. (1994), Discussion of "Sliced Inverse Regression for Dimension Reduction", by K. C. Li, *Journal of the American Statistical Association*, 86, 333–333.

Cook, R. D. (1996), "Graphics for Regressions with 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: John Wiley & Sons.

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 Nachtsheim, C. J. (1994), "Re-weighting to Achieve Elliptically Contoured Covariates in Regression", *Journal of the American Statistical Association*, 89, 592–600.

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 Weisberg, S. (1991), Discussion of "Sliced Inverse Regression for Dimension Reduction", by K. C. Li, *Journal of the American Statistical Association*, 86, 328–332.

Cook, R. D. and Yin, X. (2001), "Dimension Reduction and Visualization in Discriminant Analysis" (with discussion), *Australian and New Zealand Journal of Statistics*, 43, 147–199.

Folland, G. B. (1992), *Fourier Analysis and Its Applications*, Brooks/Cole Publishing.

Härdle, W. and Stoker, T. M. (1989), "Investigating Smooth Multiple Regression by the Method of Average Derivatives", *Journal of the American Statistical Association*, 84, 986–995.

Hirschaste, M., Juditsky, A., Polzehl, J., and Spokoiny, V. (2001), "Structure Adaptive Approach for Dimension Reduction", *The Annals of Statistics*, 29, 1537–1566.

Lee, A. J. (1990), *U-Statistics: Theory and Practice*, New York: Marcel Dekker Inc.

Li, B. and Zha, H. and Chiaromonte, F. (2005), "Contour Regression: A General Approach to Dimension Reduction", *The Annals of Statistics*, to appear.

Li, K. C. (1991), "Sliced Inverse Regression for Dimension Reduction" (with discussions), *Journal of the American Statistical Association*, 86, 316–342.

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, K. C. (1997), "Nonlinear Confounding in High-Dimensional Regression", *The Annals of Statistics*, 25, 577–612.

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

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 Cook, R. D. (2002), "Dimension reduction for the conditional kth moment in regression", *Journal of the Royal Statistical Society*, Series B, 64, 159–175.