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

Yu ZHU and Peng ZENG

In regression with a high-dimensional predictor vector, it is important to estimate the central and central mean subspaces that preserve sufficient information about the response and the mean response. 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 with SIR, SAVE, and other existing methods. The approach proposed in the article provides a novel view on sufficient dimension reduction and may lead to more powerful tools in the future.

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

1. INTRODUCTION

Suppose that Y is a univariate response and $\mathbf{X}$ is a p-dimensional vector of continuous predictors. Let $F_{Y|\mathbf{X}}$ denote the conditional distribution of Y given $\mathbf{X}$ and let $E[Y|\mathbf{X}]$ denote the mean response at $\mathbf{X}$. In full generality, the regression of Y on $\mathbf{X}$ is to infer about the conditional distribution $F_{Y|\mathbf{X}}$, often with the mean response $E[Y|\mathbf{X}]$ of primary interest. When $F_{Y|\mathbf{X}}$ or $E[Y|\mathbf{X}]$ does not admit a proper parametric form, nonparametric methods such as local polynomial smoothing are usually used for regression. Because of the curse of dimensionality, however, these methods become impractical when the dimension of $\mathbf{X}$ is high. To mitigate the curse of dimensionality, various dimension reduction techniques have been proposed in the literature, including projection pursuit regression (Friedman and Stuetzle 1981) and principal component regression (Hotelling 1957; Kendall 1957). One popular approach is to project $\mathbf{X}$ onto a lower-dimensional subspace where the regression of Y on $\mathbf{X}$ can be performed. In this article we focus on *sufficient dimension reduction* (SDR), which further requires that the projection of $\mathbf{X}$ onto the lower-dimensional subspace does not result in any loss of information about $F_{Y|\mathbf{X}}$ or $E[Y|\mathbf{X}]$.

The theory of sufficient dimension reduction originated from the seminal works of 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 let $P_{\mathcal{S}}$ be the orthogonal projection operator onto $\mathcal{S}$ in the usual inner product. $\mathcal{S}$ is called a dimension-reduction subspace if Y and $\mathbf{X}$ are independent conditioned on $P_{\mathcal{S}}\mathbf{X}$, that is,

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

where $\perp\!\!\!\perp$ means "independent of." The 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 $\mathcal{S}_{Y|\mathbf{X}}$ (Cook 1996, 1998). The dimension of $\mathcal{S}_{Y|\mathbf{X}}$ is called the *structural dimension* of the regression of Y on $\mathbf{X}$ and is denoted by $\dim(\mathcal{S}_{Y|\mathbf{X}})$. $\mathcal{S}_{Y|\mathbf{X}}$ can be considered a metaparameter that is the target of sufficient dimension reduction for $F_{Y|\mathbf{X}}$. Under mild conditions, Cook (1996, 1998) showed that the central subspace exists and is unique. Throughout this article, we assume the existence of the central subspace.

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

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

If the intersection of all mean dimension-reduction subspaces is also a mean dimension-reduction subspace, then it is considered the *central mean subspace*, denoted by $\mathcal{S}_{E[Y|\mathbf{X}]}$ (Cook and Li 2002). Similar to the central subspace, the central mean subspace exists under mild conditions, and so its existence is assumed throughout this article. $\mathcal{S}_{E[Y|\mathbf{X}]}$ is the target of sufficient dimension reduction for the mean response $E[Y|\mathbf{X}]$ and is always a subspace of the central subspace $\mathcal{S}_{Y|\mathbf{X}}$ (Cook and Li 2002). Recently, 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 $F_{Y|\mathbf{X}}$.

Various dimension-reduction methods have been proposed in the literature, some of which can be used to estimate the central subspace or the central mean subspace. For the central subspace, these include sliced inverse regression (SIR; Li 1991) and sliced average variance estimation (SAVE; Cook and Weisberg 1991); for the central mean subspace, they include principal Hessian direction (pHd; Li 1992), iterative Hessian transformation (IHT; Cook and Li 2002), the structure adaptive method (SAM; Hristache, Juditsky, Polzehl, and Spokoiny 2001), and minimum average variance estimation (MAVE; Xia, Tong, Li, and Zhu 2002). SAM and MAVE are fundamentally different from the other methods mentioned earlier in that both involve nonparametric estimation of the link function $E[Y|\mathbf{X}=\mathbf{x}]$, which may be impractical when the dimension of $\mathbf{X}$ is high. All of the other aforementioned methods avoid nonparametric estimation of the link function and target either $\mathcal{S}_{Y|\mathbf{X}}$ or $\mathcal{S}_{E[Y|\mathbf{X}]}$ directly. They usually follow a common procedure consisting of two steps. The first step is to define a $p \times p$ nonnegative definite matrix $\mathbf{M}$ called a candidate matrix (Ye

Yu Zhu is Assistant Professor, Department of Statistics, Purdue University, West Lafayette, IN 47907 (E-mail: *yuzhu@stat.purdue.edu*). Peng Zeng is Assistant Professor, Department of Mathematics and Statistics, Auburn University, Auburn, AL 36849 (E-mail: *zengpen@auburn.edu*). The authors thank the editor, the associate editor, and three anonymous referees for constructive comments and suggestions that greatly helped improve an earlier manuscript.

Journal of the American Statistical Association
December 2006, Vol. 101, No. 476, Theory and Methods
DOI 10.1198/016214506000000140

and Weiss 2003), whose columns span a subspace of $\mathcal{S}_{E[Y|\mathbf{X}]}$ or $\mathcal{S}_{Y|\mathbf{X}}$, and then propose a consistent estimate $\hat{\mathbf{M}}$ of the candidate matrix from a sample $\{(\mathbf{x}_i, y_i)\}_{1\le i\le n}$ of $(\mathbf{X}, Y)$. The second step is to obtain the spectral decomposition of $\hat{\mathbf{M}}$ and use the space spanned by the eigenvectors of $\hat{\mathbf{M}}$ corresponding to the largest q eigenvalues as the estimate of $\mathcal{S}_{E[Y|\mathbf{X}]}$ or $\mathcal{S}_{Y|\mathbf{X}}$, where q is the dimension of $\mathcal{S}_{E[Y|\mathbf{X}]}$ or $\mathcal{S}_{Y|\mathbf{X}}$. Recently, Cook and Ni (2005) proposed the minimum discrepancy method, which is more efficient than spectral decomposition for estimating $\mathcal{S}_{E[Y|\mathbf{X}]}$ or $\mathcal{S}_{Y|\mathbf{X}}$ from a given candidate matrix. For these methods to work, some distributional assumptions need to be imposed on $\mathbf{X}$. For convenience, we assume that the mean of $\mathbf{X}$ is the origin of $\mathbb{R}^p$ and the covariance matrix of $\mathbf{X}$ is the standard $p\times p$ identity matrix $\mathbf{I}_p$. Then SIR and IHT require that $\mathbf{X}$ satisfies the linearity condition, $E[\mathbf{X}|P_{\mathcal{S}_{Y|\mathbf{X}}}\mathbf{X}] = P_{\mathcal{S}_{Y|\mathbf{X}}}\mathbf{X}$, and SAVE and pHd require an additional condition called the constant variance condition, $\operatorname{cov}[\mathbf{X}|P_{\mathcal{S}_{Y|\mathbf{X}}}\mathbf{X}] = \mathbf{Q}_{\mathcal{S}_{Y|\mathbf{X}}}$, where $\mathbf{Q}_{\mathcal{S}_{Y|\mathbf{X}}} = \mathbf{I}_p - P_{\mathcal{S}_{Y|\mathbf{X}}}$. These conditions are satisfied when the distribution of $\mathbf{X}$ is multivariate normal. (For a detailed discussion of the conditions, see Cook 1998.)

Although SIR, SAVE, pHd, and IHT work well in practice, there has not been much study on when they can exhaustively recover the central subspace or the central mean subspace. It is known that SIR fails to capture directions along which Y is symmetric about. For example, if $Y = (\boldsymbol{\beta}^\tau\mathbf{X})^2 + \varepsilon$, where $\boldsymbol{\beta}$ is a p-dimensional vector, τ denotes transpose, $\mathbf{X}$ follows $\mathrm{N}(\mathbf{0}, \mathbf{I}_p)$, and ε is a random error independent of $\mathbf{X}$, then SIR will miss $\boldsymbol{\beta}$. A sufficient condition for SAVE to exhaustively recover $\mathcal{S}_{Y|\mathbf{X}}$ is that the conditional distribution of $\mathbf{X}$ 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 $F_{Y|\mathbf{X}}$ or $E[Y|\mathbf{X}]$. 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 $\mathbf{X}$ follows an elliptically contoured distribution and also requires conditions that involve $\mathbf{X}$, Y, and both vectors in $\mathcal{S}_{Y|\mathbf{X}}$ and $\mathcal{S}_{Y|\mathbf{X}}^{\perp}$ (see assumptions 2.1 and 4.1 and thm. 4.2 in Li et al. 2005).

This article represents another effort to derive methods that can fully recover the central mean subspace and the central subspace under various conditions. The primary tool that we use is the Fourier transform. At the population level, we have derived two candidate matrices, $\mathbf{M}_{\mathrm{FM}}$ and $\mathbf{M}_{\mathrm{FC}}$, whose column spaces are identical to the central mean subspace and the central subspace. Given a sample of $(\mathbf{X}, Y)$, if consistent estimates of $\mathbf{M}_{\mathrm{FM}}$ and $\mathbf{M}_{\mathrm{FC}}$ 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 that we have regarding the distribution of $\mathbf{X}$. Due to space limitations, in this article we fully implement our methods only for the case where $\mathbf{X}$ is normally distributed. The implementation of our method under more general conditions is briefly described herein and is currently under further investigation, and we will report the results in future work.

To facilitate our approach, we need to modify the model assumptions as follows. First, we assume that the joint distribution of $(\mathbf{X}, Y)$, the conditional distributions of $\mathbf{X}|Y$ and $Y|\mathbf{X}$, and the marginal distributions of $\mathbf{X}$ and Y admit densities, which are denoted by $f_{\mathbf{X},Y}(\mathbf{x}, y)$, $f_{Y|\mathbf{X}}(y|\mathbf{x})$, $f_{\mathbf{X}|Y}(\mathbf{x}|y)$, $f_{\mathbf{X}}(\mathbf{x})$, and $f_Y(y)$. Let $\mathbf{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_q)$ be a $p\times q$ matrix with its columns forming a basis for $\mathcal{S}_{Y|\mathbf{X}}$. Then (1) can be restated in terms of conditional distributions as $F_{Y|\mathbf{X}} = F_{Y|\mathbf{B}^\tau\mathbf{X}}$ or in terms of density as

$$f_{Y|\mathbf{X}}(y|\mathbf{x}) = f_{Y|\mathbf{B}^\tau\mathbf{X}}(y|\mathbf{B}^\tau\mathbf{x}) = h(y; \boldsymbol{\beta}_1^\tau\mathbf{x}, \boldsymbol{\beta}_2^\tau\mathbf{x}, \ldots, \boldsymbol{\beta}_q^\tau\mathbf{x}), \qquad (3)$$

where $h(y; u_1, \ldots, u_q)$ is a $(q+1)$-variate function. Similarly, let $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_q$ be a basis of $\mathcal{S}_{E[Y|\mathbf{X}]}$; then (2) is equivalent to

$$E[Y|\mathbf{X} = \mathbf{x}] = g(\boldsymbol{\alpha}_1^\tau\mathbf{x}, \boldsymbol{\alpha}_2^\tau\mathbf{x}, \ldots, \boldsymbol{\alpha}_q^\tau\mathbf{x}), \qquad (4)$$

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

2. CENTRAL MEAN SUBSPACE

In this section we propose a candidate matrix $\mathbf{M}_{\mathrm{FM}}$ whose column space is exactly the central mean subspace $\mathcal{S}_{E[Y|\mathbf{X}]}$. We follow a commonly used idea for deriving candidate matrices in the literature: identify vectors that belong to $\mathcal{S}_{E[Y|\mathbf{X}]}$ and combine them to generate the candidate matrix. The major tool that we use is the Fourier transform.

Let $m(\mathbf{x}) = E[Y|\mathbf{X} = \mathbf{x}]$. From (4), $m(\mathbf{x}) = g(\mathbf{u})$, where $\mathbf{u} = \mathbf{A}^\tau\mathbf{x}$ and $\mathbf{A} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_q)$, whose columns form a basis of $\mathcal{S}_{E[Y|\mathbf{X}]}$. Let $\frac{\partial}{\partial\mathbf{x}} = (\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \ldots, \frac{\partial}{\partial x_p})^\tau$ denote the gradient operator. By the chain rule of differentiation,

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

Thus the gradient of $m(\mathbf{x})$ at any fixed $\mathbf{x}$ is a linear combination of $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_q$; therefore, it is in $\mathcal{S}_{E[Y|\mathbf{X}]}$. Let $\operatorname{supp}(\mathbf{X}) = \{\mathbf{x}\in\mathbb{R}^p : f_{\mathbf{X}}(\mathbf{x}) > 0\}$ be the support of $\mathbf{X}$. The collection of all of the gradients of $m(\mathbf{x})$ over $\mathbf{x}\in\operatorname{supp}(\mathbf{X})$ spans the central mean subspace $\mathcal{S}_{E[Y|\mathbf{X}]}$, as does the collection of all the gradients of $m(\mathbf{x})$ weighted by $f_{\mathbf{X}}(\mathbf{x})$, that is,

$$\begin{aligned}\mathcal{S}_{E[Y|\mathbf{X}]} &= \operatorname{span}\left\{\frac{\partial}{\partial\mathbf{x}}m(\mathbf{x}), \mathbf{x}\in\operatorname{supp}(\mathbf{X})\right\}\\ &= \operatorname{span}\left\{\left(\frac{\partial}{\partial\mathbf{x}}m(\mathbf{x})\right)f_{\mathbf{X}}(\mathbf{x}), \mathbf{x}\in\mathbb{R}^p\right\}. \qquad (5)\end{aligned}$$

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

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

$$\boldsymbol{\psi}(\boldsymbol{\omega}) = \int \exp\{\imath\boldsymbol{\omega}^\tau \mathbf{x}\}\left(\frac{\partial}{\partial\mathbf{x}}m(\mathbf{x})\right)f_{\mathbf{X}}(\mathbf{x})\,d\mathbf{x}. \tag{6}$$

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

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

$$\left(\frac{\partial}{\partial\mathbf{x}}m(\mathbf{x})\right)f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-p}\int \exp\{-\imath\mathbf{x}^\tau\boldsymbol{\omega}\}\boldsymbol{\psi}(\boldsymbol{\omega})\,d\boldsymbol{\omega}. \tag{7}$$

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

Proposition 1.

1. The central mean subspace is spanned by $\mathbf{a}(\boldsymbol{\omega})$ and $\mathbf{b}(\boldsymbol{\omega})$, that is, $\mathcal{S}_{E[Y|\mathbf{X}]} = \operatorname{span}\{\mathbf{a}(\boldsymbol{\omega}), \mathbf{b}(\boldsymbol{\omega}) : \boldsymbol{\omega} \in \mathbb{R}^p\}$.
2. Suppose that $\log f_{\mathbf{X}}(\mathbf{x})$ is differentiable and $m(\mathbf{x})f_{\mathbf{X}}(\mathbf{x})$ goes to 0 as $\|\mathbf{x}\| \to \infty$, then
$$\boldsymbol{\psi}(\boldsymbol{\omega}) = -E_{(\mathbf{X},Y)}\big[(\imath\boldsymbol{\omega} + \mathbf{G}(\mathbf{X}))Y\exp\{\imath\boldsymbol{\omega}^\tau\mathbf{X}\}\big], \tag{8}$$
where $\mathbf{G}(\mathbf{x}) = \frac{\partial}{\partial\mathbf{x}}\log f_{\mathbf{X}}(\mathbf{x})$.
3. If $(\frac{\partial}{\partial x_i}m(\mathbf{x}))f_{\mathbf{X}}(\mathbf{x})$ is absolutely integrable for $1 \le i \le p$, then $\boldsymbol{\psi}(\boldsymbol{\omega}) \to \mathbf{0}$ as $\|\boldsymbol{\omega}\| \to \infty$.
4. If $(\frac{\partial}{\partial x_i}m(\mathbf{x}))f_{\mathbf{X}}(\mathbf{x})$ is squared integrable for $1 \le i \le p$, then
$$\int\left(\frac{\partial}{\partial\mathbf{x}}m(\mathbf{x})\right)\left(\frac{\partial}{\partial\mathbf{x}}m(\mathbf{x})\right)^\tau f_{\mathbf{X}}(\mathbf{x})^2\,d\mathbf{x} = (2\pi)^{-p}\int \boldsymbol{\psi}(\boldsymbol{\omega})\bar{\boldsymbol{\psi}}(\boldsymbol{\omega})^\tau\,d\boldsymbol{\omega}, \tag{9}$$
where $\bar{\boldsymbol{\psi}}(\boldsymbol{\omega})$ is the conjugate of $\boldsymbol{\psi}(\boldsymbol{\omega})$.

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

The third property is essentially the Riemann–Lebesgue lemma for the Fourier transform (Folland 1992, pp. 217, 243). It indicates that when the density-weighted gradient is absolutely integrable, $\boldsymbol{\psi}(\boldsymbol{\omega})$ decays to $\mathbf{0}$ as the norm of $\boldsymbol{\omega}$ goes to infinity. The fourth property is a result from applying the Plancheral theorem for the Fourier transform (Folland 1992, pp. 222, 224) to $(\frac{\partial}{\partial\mathbf{x}}m(\mathbf{x}))f_{\mathbf{X}}(\mathbf{x})$ and $\boldsymbol{\psi}(\boldsymbol{\omega})$, and it establishes the connection between the expected outer product of the density-weighted gradient of $m(\mathbf{x})$ (i.e., $E[(\frac{\partial}{\partial\mathbf{x}}m(\mathbf{X}))(\frac{\partial}{\partial\mathbf{x}}m(\mathbf{X}))^\tau f_{\mathbf{X}}(\mathbf{X})]$), and the integral of the outer product of $\boldsymbol{\psi}(\boldsymbol{\omega})$. Let $\mathbf{M}^*_{\mathrm{FM}} = (2\pi)^{-p}\int \boldsymbol{\psi}(\boldsymbol{\omega})\bar{\boldsymbol{\psi}}(\boldsymbol{\omega})^\tau\,d\boldsymbol{\omega}$. Then the column space of $\mathbf{M}^*_{\mathrm{FM}}$, $\mathcal{S}(\mathbf{M}^*_{\mathrm{FM}})$, is exactly equal to the central mean subspace, as stated in the following proposition.

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

Intuitively, $\mathbf{M}^*_{\mathrm{FM}}$ can be considered the sum of the outer product of $\boldsymbol{\psi}(\boldsymbol{\omega})$ over all $\boldsymbol{\omega}$. Because of (7), $\boldsymbol{\psi}(\boldsymbol{\omega})$ can be considered the vector of coefficients of $\exp\{\imath\mathbf{x}^\tau\boldsymbol{\omega}\}$ in the representation of $\frac{\partial}{\partial\mathbf{x}}m(\mathbf{x})f_{\mathbf{X}}(\mathbf{x})$. For different $\boldsymbol{\omega}$'s, the $\exp\{\imath\mathbf{x}^\tau\boldsymbol{\omega}\}$ are basic oscillatory functions with different frequencies. So $\boldsymbol{\psi}(\boldsymbol{\omega})$ with small $\|\boldsymbol{\omega}\|$ captures the patterns of $\frac{\partial}{\partial\mathbf{x}}m(\mathbf{x})f_{\mathbf{X}}(\mathbf{x})$ with low frequencies, whereas $\boldsymbol{\psi}(\boldsymbol{\omega})$ with large $\|\boldsymbol{\omega}\|$ captures the patterns of $\frac{\partial}{\partial\mathbf{x}}m(\mathbf{x})f_{\mathbf{X}}(\mathbf{x})$ with high frequencies. According to the third property of $\boldsymbol{\psi}(\boldsymbol{\omega})$, $\boldsymbol{\psi}(\boldsymbol{\omega})$ goes to $\mathbf{0}$ when $\|\boldsymbol{\omega}\|$ goes to infinity. Therefore, when patterns with various frequencies are of different interest, $\boldsymbol{\psi}(\boldsymbol{\omega})$ of different $\boldsymbol{\omega}$'s should be treated differently. This can be realized by assigning different weights to $\boldsymbol{\omega}$ when combining the outer product of $\boldsymbol{\psi}(\boldsymbol{\omega})$. We use $K(\boldsymbol{\omega})$ to denote the weight function and generate a more flexible candidate matrix for the central mean subspace as

$$\begin{aligned}\mathbf{M}_{\mathrm{FM}} &= \operatorname{Re}\left(\int \boldsymbol{\psi}(\boldsymbol{\omega})\bar{\boldsymbol{\psi}}(\boldsymbol{\omega})^\tau K(\boldsymbol{\omega})\,d\boldsymbol{\omega}\right)\\ &= \int [\mathbf{a}(\boldsymbol{\omega})\mathbf{a}(\boldsymbol{\omega})^\tau + \mathbf{b}(\boldsymbol{\omega})\mathbf{b}(\boldsymbol{\omega})^\tau]K(\boldsymbol{\omega})\,d\boldsymbol{\omega}, \end{aligned}\tag{10}$$

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

Proposition 3. If $K(\boldsymbol{\omega})$ is a positive weight function on $\mathbb{R}^p$, then $\mathbf{M}_{\mathrm{FM}}$ is a nonnegative definite matrix and $\mathcal{S}(\mathbf{M}_{\mathrm{FM}}) = \mathcal{S}_{E[Y|\mathbf{X}]}$.

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

$$\mathbf{M}_{\mathrm{FM}} = E_{(\mathbf{U}_1,V_1),(\mathbf{U}_2,V_2)}\big[\mathbf{J}_{\mathrm{FM}}((\mathbf{U}_1,V_1),(\mathbf{U}_2,V_2))\big], \tag{11}$$

where

$$\begin{aligned}\mathbf{J}_{\mathrm{FM}}&((\mathbf{U}_1,V_1),(\mathbf{U}_2,V_2))\\ &= V_1V_2e^{-\sigma_{\mathrm{W}}^2/2\|\mathbf{U}_{12}\|^2}\\ &\quad\times\big[\sigma_{\mathrm{W}}^2\mathbf{I}_p + (\mathbf{G}(\mathbf{U}_1)-\sigma_{\mathrm{W}}^2\mathbf{U}_{12})(\mathbf{G}(\mathbf{U}_2)+\sigma_{\mathrm{W}}^2\mathbf{U}_{12})^\tau\big],\end{aligned}$$

$(\mathbf{U}_1, V_1)$ and $(\mathbf{U}_2, V_2)$ are independent and identically distributed as $(\mathbf{X}, Y)$, $\mathbf{I}_p$ is the $p \times p$ identity matrix, and $\mathbf{U}_{12} = \mathbf{U}_1 - \mathbf{U}_2$.

3. CENTRAL SUBSPACE

As discussed in Section 1, the central mean subspace can only capture the information in $\mathbf{X}$ regarding the mean response $E[Y|\mathbf{X}]$. In applications where the entire conditional distribution $F_{Y|\mathbf{X}}$ is of interest, sufficient dimension reduction should aim at the central subspace $\mathcal{S}_{Y|\mathbf{X}}$. This section focuses on the derivation of the candidate matrix $\mathbf{M}_{\mathrm{FC}}$ for $\mathcal{S}_{Y|\mathbf{X}}$. We again use the Fourier transform, as well as other similar ideas from Section 2.

First, we establish a connection between $\mathcal{S}_{Y|\mathbf{X}}$ and a family of central mean subspaces. As noted in Section 1, the central mean subspace $\mathcal{S}_{E[Y|\mathbf{X}]}$ is always a subspace of $\mathcal{S}_{Y|\mathbf{X}}$. 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 $\mathcal{S}_{E[T(Y)|\mathbf{X}]}$, is also a subspace of $\mathcal{S}_{Y|\mathbf{X}}$. For two different transformations $T_1(Y)$ and $T_2(Y)$, their corresponding central mean subspaces $\mathcal{S}_{E[T_1(Y)|\mathbf{X}]}$ and $\mathcal{S}_{E[T_2(Y)|\mathbf{X}]}$ are not necessarily identical and may cover different parts of $\mathcal{S}_{Y|\mathbf{X}}$. This provides a possibility to recover the entire central subspace by collecting the central mean subspace of $T(Y)$ over all of the possible transformations, that is,

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

where

$$\begin{aligned}\sum_T &\mathcal{S}_{E[T(Y)|\mathbf{X}]}\\ &= \{u: \text{There exist a finite number of transformations}\\ &\qquad T_1,\dots,T_k, \text{ such that } u = u_1+\cdots+u_k \text{ and}\\ &\qquad u_i \in \mathcal{S}_{E[T_i(Y)|\mathbf{X}]} \text{ for } 1\le i\le k\}.\end{aligned}$$

The foregoing equation is indeed true and is implied by Proposition 4. In fact, it is not necessary to use all of the possible transformations in (12); a family of properly chosen transformations is sufficient to serve that 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 $\mathbf{X} = \mathbf{x}$ is

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

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

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

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

Because $\{T(\cdot,t): t\in\mathbb{R}\}$ is a subset of all the possible transformations, Proposition 4 implies (12).

When defining the central kth-moment subspace, Yin and Cook (2002) considered the power transformations of Y, which are Y^l with $1 \le l \le k$. Although the transformations that we consider can be considered an extension of the power transformations, they lead to entirely different methods for recovering the central subspace. Proposition 4 suggests that, to recover the central subspace $\mathcal{S}_{Y|\mathbf{X}}$, 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 $\mathcal{S}_{E[Y|\mathbf{X}]}$ can be directly generalized for the central mean subspace $\mathcal{S}_{E[T(Y,t)|\mathbf{X}]}$ with Y replaced by $\exp\{\imath tY\}$. Hence we can combine the candidate matrices for $\mathcal{S}_{E[T(Y,t)|\mathbf{X}]}$ to generate a candidate matrix for $\mathcal{S}_{Y|\mathbf{X}}$. The idea of combining candidate matrices to generate new ones was originally mentioned by Li (1991) and was further investigated by Ye and Weiss (2003). In what follows, we start with applying the Fourier transform to the gradient of $m(\mathbf{x},t)$ as in Section 2 and materialize the foregoing idea to derive a candidate matrix $\mathbf{M}_{\mathrm{FC}}$ for $\mathcal{S}_{Y|\mathbf{X}}$.

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

$$\begin{aligned}\mathcal{S}_{E[T(Y,t)|\mathbf{X}]} &= \mathrm{span}\left\{\frac{\partial}{\partial\mathbf{x}}m(\mathbf{x},t), \mathbf{x}\in\mathrm{supp}(\mathbf{X})\right\}\\ &= \mathrm{span}\left\{\left(\frac{\partial}{\partial\mathbf{x}}m(\mathbf{x},t)\right)f_{\mathbf{X}}(\mathbf{x}), \mathbf{x}\in\mathbb{R}^p\right\},\end{aligned}$$

where 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

$$\begin{aligned}\mathcal{S}_{Y|\mathbf{X}} &= \mathrm{span}\left\{\frac{\partial}{\partial\mathbf{x}}m(\mathbf{x},t), \mathbf{x}\in\mathrm{supp}(\mathbf{X}), t\in\mathbb{R}\right\}\\ &= \mathrm{span}\left\{\left(\frac{\partial}{\partial\mathbf{x}}m(\mathbf{x},t)\right)f_{\mathbf{X}}(\mathbf{x}), \mathbf{x}\in\mathbb{R}^p, t\in\mathbb{R}\right\}.\end{aligned} \tag{13}$$

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

$$\boldsymbol{\phi}(\boldsymbol{\omega},t) = \int \exp\{\imath\boldsymbol{\omega}^\tau\mathbf{x}\}\left(\frac{\partial}{\partial\mathbf{x}}m(\mathbf{x},t)\right)f_{\mathbf{X}}(\mathbf{x})\,d\mathbf{x}. \tag{14}$$

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

Proposition 5. For each fixed $t \in \mathbb{R}$, the following assertions hold:

1. Both $\mathbf{a}(\boldsymbol{\omega}, t)$ and $\mathbf{b}(\boldsymbol{\omega}, t)$ are vectors in $\mathcal{S}_{E[T(Y,t)|\mathbf{X}]}$; furthermore,
$$\mathcal{S}_{Y|\mathbf{X}} = \operatorname{span}\{\mathbf{a}(\boldsymbol{\omega}, t), \mathbf{b}(\boldsymbol{\omega}, t) : \boldsymbol{\omega} \in \mathbb{R}^p, t \in \mathbb{R}\}. \qquad (15)$$
2. Suppose that $\log f_{\mathbf{X}}(\mathbf{x})$ is differentiable and that $m(\mathbf{x}, t) f_{\mathbf{X}}(\mathbf{x})$ goes to 0 when $\|\mathbf{x}\| \to \infty$; then
$$\boldsymbol{\phi}(\boldsymbol{\omega}, t) = -E_{(\mathbf{X},Y)}\big[(\imath\boldsymbol{\omega} + \mathbf{G}(\mathbf{X}))\exp\{\imath\boldsymbol{\omega}^{\tau}\mathbf{X} + \imath tY\}\big]. \qquad (16)$$
3. If $(\frac{\partial}{\partial x_i} m(\mathbf{x}, t)) f_{\mathbf{X}}(\mathbf{x})$ is absolutely integrable for $1 \le i \le p$, then $\boldsymbol{\phi}(\boldsymbol{\omega}, t) \to \mathbf{0}$ as $\|\boldsymbol{\omega}\| \to \infty$.
4. If $(\frac{\partial}{\partial x_i} m(\mathbf{x}, t)) f_{\mathbf{X}}(\mathbf{x})$ is squared integrable for $1 \le i \le p$, then
$$\int \left(\frac{\partial}{\partial \mathbf{x}} m(\mathbf{x}, t)\right)\left(\frac{\partial}{\partial \mathbf{x}} \bar{m}(\mathbf{x}, t)\right)^{\tau} f_{\mathbf{X}}(\mathbf{x})^2 \, d\mathbf{x} = (2\pi)^{-p} \int \boldsymbol{\phi}(\boldsymbol{\omega}, t)\bar{\boldsymbol{\phi}}(\boldsymbol{\omega}, t)^{\tau} \, d\boldsymbol{\omega}, \qquad (17)$$
where $\bar{m}(\mathbf{x}, t)$ and $\bar{\boldsymbol{\phi}}(\boldsymbol{\omega}, t)$ are the conjugates of $m(\mathbf{x}, t)$ and $\boldsymbol{\phi}(\boldsymbol{\omega}, t)$.

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

$$\begin{aligned}\mathbf{M}_{\mathrm{FC}} &= \operatorname{Re}\left(\iint \boldsymbol{\phi}(\boldsymbol{\omega}, t)\bar{\boldsymbol{\phi}}(\boldsymbol{\omega}, t)^{\tau} K(\boldsymbol{\omega})k(t)\, d\boldsymbol{\omega}\, dt\right) \\ &= \iint [\mathbf{a}(\boldsymbol{\omega}, t)\mathbf{a}(\boldsymbol{\omega}, t)^{\tau} + \mathbf{b}(\boldsymbol{\omega}, t)\mathbf{b}(\boldsymbol{\omega}, t)^{\tau}] \\ &\qquad \times K(\boldsymbol{\omega})k(t)\, d\boldsymbol{\omega}\, dt. \end{aligned} \qquad (18)$$

Proposition 6. If $K(\boldsymbol{\omega})$ is a positive weight function on $\mathbb{R}^p$ and $k(t)$ is a positive weight function on $\mathbb{R}$, then $\mathbf{M}_{\mathrm{FC}}$ is a nonnegative definite matrix, and $\mathcal{S}(\mathbf{M}_{\mathrm{FC}}) = \mathcal{S}_{Y|\mathbf{X}}$.

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

$$\mathbf{M}_{\mathrm{FC}} = E_{(\mathbf{U}_1,V_1),(\mathbf{U}_2,V_2)}\big[\mathbf{J}_{\mathrm{FC}}((\mathbf{U}_1, V_1), (\mathbf{U}_2, V_2))\big], \qquad (19)$$

where

$$\begin{aligned}\mathbf{J}_{\mathrm{FC}}&((\mathbf{U}_1, V_1), (\mathbf{U}_2, V_2)) \\ &= \exp\left\{-\frac{\sigma_{\mathrm{W}}^2}{2}\|\mathbf{U}_{12}\|^2 - \frac{\sigma_{\mathrm{T}}^2}{2}V_{12}^2\right\} \\ &\qquad \times \big[\sigma_{\mathrm{W}}^2\mathbf{I}_p + (\mathbf{G}(\mathbf{U}_1) - \sigma_{\mathrm{W}}^2\mathbf{U}_{12})(\mathbf{G}(\mathbf{U}_2) + \sigma_{\mathrm{W}}^2\mathbf{U}_{12})^{\tau}\big], \end{aligned} \qquad (20)$$

$(\mathbf{U}_1, V_1)$ and $(\mathbf{U}_2, V_2)$ are independent and identically distributed as $(\mathbf{X}, Y)$, $\mathbf{I}_p$ is the identity matrix, $\mathbf{U}_{12} = \mathbf{U}_1 - \mathbf{U}_2$, and $V_{12} = V_1 - V_2$.

The derivation of $\mathbf{M}_{\mathrm{FC}}$ can also be understood from the perspective of inverse regression used in SIR, where the candidate matrix is defined by the first moment of the conditional distribution of $\mathbf{X}$ given Y. Next we present a connection between $\boldsymbol{\phi}(\boldsymbol{\omega}, t)$, which is used to define $\mathbf{M}_{\mathrm{FC}}$, and the conditional distribution of $\mathbf{X}$ given Y. Define

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

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

Proposition 7. Suppose for any fixed y, $f_{\mathbf{X},Y}(\mathbf{x}, y)$ goes to 0 as $\|\mathbf{x}\| \to \infty$. Then the following hold:

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

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

Proposition 8. Suppose that $\mathbf{X}$ follows the normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}_p$. If $K(\boldsymbol{\omega})$ is chosen to be a point mass at $\boldsymbol{\omega} = \mathbf{0}$, then $\mathcal{S}(\mathbf{M}_{\mathrm{FC}}) = \mathcal{S}(\mathbf{M}_{\mathrm{SIR}})$.

Proposition 8 indicates that the column spaces of $\mathbf{M}_{\mathrm{FC}}$ and $\mathbf{M}_{\mathrm{SIR}}$ coincide when $\mathbf{X}$ follows the standard normal distribution and the weight function $K(\boldsymbol{\omega})$ is degenerate at $\boldsymbol{\omega} = \mathbf{0}$. Hence $\mathbf{M}_{\mathrm{SIR}}$ could be considered a special case of $\mathbf{M}_{\mathrm{FC}}$ under the normality assumption. Although $\mathcal{S}(\mathbf{M}_{\mathrm{FC}})$ and $\mathcal{S}(\mathbf{M}_{\mathrm{SIR}})$ are the same, the matrices are generally different from each other. It is known that SIR fails to capture the directions along which Y is symmetric about, as does $\mathbf{M}_{\mathrm{FC}}$ with $\sigma_{\mathrm{W}}^2 = 0$. In general, we do not use a degenerate weight function for $\boldsymbol{\omega}$. 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 $\mathbf{M}_{\mathrm{FC}}$ and $\mathbf{M}_{\mathrm{FM}}$ 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 discuss their selection in the next section.

Let $\{(\mathbf{x}_i, y_i)\}_{1\le i\le n}$ be a random sample of $(\mathbf{X}, Y)$. Without loss of generality, we assume that $E[\mathbf{X}] = \mathbf{0}$ and $\operatorname{cov}[\mathbf{X}] = \mathbf{I}_p$. We first consider $\mathbf{M}_{\text{FC}}$. According to (19), $\mathbf{M}_{\text{FC}}$ is the expectation of $\mathbf{J}_{\text{FC}}$ over $(\mathbf{U}_1, V_1)$ and $(\mathbf{U}_2, V_2)$ that are independent and identically distributed as $(\mathbf{X}, Y)$. Let $F(\mathbf{x}, y)$ be the cumulative distribution function of $(\mathbf{X}, Y)$. Then $\mathbf{M}_{\text{FC}}$ can be expressed as

$$\mathbf{M}_{\text{FC}} = \iint \mathbf{J}_{\text{FC}}((\mathbf{u}_1, v_1), (\mathbf{u}_2, v_2))\, dF(\mathbf{u}_1, v_1)\, dF(\mathbf{u}_2, v_2). \tag{21}$$

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

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

where $I_{[\cdot]}$ is the indicator function. Therefore, a proper estimate of $\mathbf{M}_{\text{FC}}$ is derived by replacing $F(\mathbf{u}_1, v_1)$ and $F(\mathbf{u}_2, v_2)$ in (21) with $F_n(\mathbf{u}_1, v_1)$ and $F_n(\mathbf{u}_2, v_2)$, which is

$$\hat{\mathbf{M}}_{\text{FC}} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{J}_{\text{FC}}((\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j)). \tag{22}$$

The explicit expression of $\mathbf{J}_{\text{FC}}((\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j))$ is given in (20). Note that to make $\hat{\mathbf{M}}_{\text{FC}}$ a legitimate estimate, we need to estimate $\mathbf{G}(\mathbf{x}_i) = \frac{\partial}{\partial \mathbf{x}} \log f_{\mathbf{X}}(\mathbf{x}_i)$ for $1 \le i \le n$.

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

In the rest of this article, we focus on the case where $\mathbf{X}$ 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 herein can still be applied. (See Brillinger 1994 for the resampling method and Cook and Nachtsheim 1994 for the Voronoi weighting method.)

Assume that $\mathbf{X}$ follows the multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}_p$. Then $\mathbf{G}(\mathbf{x}) = -\mathbf{x}$. For clarity, we use $\mathbf{J}_{\text{FCN}}$, $\mathbf{M}_{\text{FCN}}$, and $\hat{\mathbf{M}}_{\text{FCN}}$ to denote $\mathbf{J}_{\text{FC}}$, $\mathbf{M}_{\text{FC}}$, and $\hat{\mathbf{M}}_{\text{FC}}$ under the normality assumption. Hence

$$\mathbf{M}_{\text{FCN}} = E_{(\mathbf{U}_1, V_1), (\mathbf{U}_2, V_2)}\big[\mathbf{J}_{\text{FCN}}((\mathbf{U}_1, V_1), (\mathbf{U}_2, V_2))\big] \tag{23}$$

and

$$\hat{\mathbf{M}}_{\text{FCN}} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{J}_{\text{FCN}}((\mathbf{u}_i, v_i), (\mathbf{u}_j, v_j)), \tag{24}$$

where

$$\begin{aligned}
&\mathbf{J}_{\text{FCN}}((\mathbf{u}_1, v_1), (\mathbf{u}_2, v_2)) \\
&\quad = \exp\left\{-\frac{\sigma_{\text{W}}^2}{2}\|\mathbf{u}_{12}\|^2 - \frac{\sigma_{\text{T}}^2}{2} v_{12}^2\right\} \\
&\qquad \times \big[\sigma_{\text{W}}^2 \mathbf{I}_p + (\mathbf{u}_1 + \sigma_{\text{W}}^2 \mathbf{u}_{12})(\mathbf{u}_2 - \sigma_{\text{W}}^2 \mathbf{u}_{12})^\tau\big].
\end{aligned} \tag{25}$$

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

Theorem 1. Suppose that $\mathbf{X}$ follows the standard multivariate normal distribution and that the covariance matrix of $\operatorname{vec}(\mathbf{J}_{\text{FCN}}((\mathbf{U}_1, V_1), (\mathbf{U}_2, V_2)))$ exists. As $n \to \infty$,

$$\hat{\mathbf{M}}_{\text{FCN}} = \mathbf{M}_{\text{FCN}} + \frac{1}{n}\sum_{i=1}^{n}\big(\mathbf{J}_{\text{FCN}}^{(1)}(\mathbf{x}_i, y_i) - 2\mathbf{M}_{\text{FCN}}\big) + o_p\big(n^{-1/2}\big),$$

where

$$\begin{aligned}
\mathbf{J}_{\text{FCN}}^{(1)}(\mathbf{x}, y) = E_{(\mathbf{U}_2, V_2)}\big[&\mathbf{J}_{\text{FCN}}((\mathbf{x}, y), (\mathbf{U}_2, V_2)) \\
&+ \mathbf{J}_{\text{FCN}}((\mathbf{x}, y), (\mathbf{U}_2, V_2))^\tau\big].
\end{aligned}$$

Let $\boldsymbol{\Sigma}_{\text{FCN}}$ be the covariance matrix of $\operatorname{vec}(\mathbf{J}_{\text{FCN}}^{(1)}(\mathbf{X}, Y))$. Then

$$\sqrt{n}\big(\operatorname{vec}(\hat{\mathbf{M}}_{\text{FCN}}) - \operatorname{vec}(\mathbf{M}_{\text{FCN}})\big) \xrightarrow{\mathcal{L}} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\text{FCN}}), \quad \text{as } n \to \infty.$$

In Theorem 1, vec is an operator that transforms a matrix to a vector by stacking up all of its columns. For instance, if $\mathbf{M} = (\mathbf{m}_1, \ldots, \mathbf{m}_k)$ is a $p \times k$ matrix and the $\mathbf{m}_i$'s are the column vectors, then $\operatorname{vec}(\mathbf{M}) = (\mathbf{m}_1^\tau, \ldots, \mathbf{m}_k^\tau)^\tau$ is a $kp \times 1$ vector. Theorem 1 asserts that $\hat{\mathbf{M}}_{\text{FCN}}$ converges to $\mathbf{M}_{\text{FCN}}$ at the rate of $\sqrt{n}$, which implies that the eigenvalues and eigenvectors of $\hat{\mathbf{M}}_{\text{FCN}}$ converge to those of $\mathbf{M}_{\text{FCN}}$ at the same rate.

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

Proposition 9. Suppose that $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q)$ is an orthonormal basis of $\mathcal{S}_{Y|\mathbf{X}}$ and that $\tilde{\mathbf{B}} = (\boldsymbol{\beta}_{q+1}, \ldots, \boldsymbol{\beta}_p)$ is an orthonormal basis of the complementary space of $\mathcal{S}_{Y|\mathbf{X}}$ in $\mathbb{R}^p$. If $\mathbf{B}^\tau\mathbf{X}$ and $\tilde{\mathbf{B}}^\tau\mathbf{X}$ are independent of each other and $\tilde{\mathbf{B}}^\tau\mathbf{X}$ follows the standard normal distribution, then $\mathcal{S}(\mathbf{M}_{\text{FCN}}) \subseteq \mathcal{S}_{Y|\mathbf{X}}$.

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

For the central mean subspace $\mathcal{S}_{E[Y|\mathbf{X}]}$ and its candidate matrix $\mathbf{M}_{\text{FM}}$, similar discussions can be used to derive the estimates of $\mathbf{M}_{\text{FM}}$ under various conditions on $\mathbf{X}$. In what follows, we report only the results under the normality assumption on $\mathbf{X}$. Again, we use $\mathbf{J}_{\text{FMN}}$, $\mathbf{M}_{\text{FMN}}$, and $\hat{\mathbf{M}}_{\text{FMN}}$ for $\mathbf{J}_{\text{FM}}$,

$\mathbf{M}_{\text{FM}}$, and $\hat{\mathbf{M}}_{\text{FM}}$ under the normality condition. Recall that $\mathbf{G}(\mathbf{x}) = -\mathbf{x}$; therefore,

$$\mathbf{M}_{\text{FMN}} = E_{(\mathbf{U}_1,V_1),(\mathbf{U}_2,V_2)}\big[\mathbf{J}_{\text{FMN}}((\mathbf{U}_1, V_1), (\mathbf{U}_2, V_2))\big], \quad (26)$$

and the estimate of $\mathbf{M}_{\text{FMN}}$ is

$$\hat{\mathbf{M}}_{\text{FMN}} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{J}_{\text{FMN}}((\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j)), \quad (27)$$

where

$$\begin{aligned}\mathbf{J}_{\text{FMN}}&((\mathbf{u}_1, v_1), (\mathbf{u}_2, v_2))\\ &= v_1 v_2 \exp\left\{-\frac{\sigma_{\text{W}}^2}{2}\|\mathbf{u}_{12}\|^2\right\}\\ &\quad\times \big[\sigma_{\text{W}}^2\mathbf{I}_p + (\mathbf{u}_1 + \sigma_{\text{W}}^2\mathbf{u}_{12})(\mathbf{u}_2 - \sigma_{\text{W}}^2\mathbf{u}_{12})^\tau\big]. \quad (28)\end{aligned}$$

The asymptotic normality of $\hat{\mathbf{M}}_{\text{FMN}}$ is established in the following theorem.

Theorem 2. Suppose that $\mathbf{X}$ follows the standard multivariate normal distribution and that the covariance matrix of $\text{vec}(\mathbf{J}_{\text{FMN}}((\mathbf{U}_1, V_1), (\mathbf{U}_2, V_2)))$ exists. As $n \to \infty$,

$$\hat{\mathbf{M}}_{\text{FMN}} = \mathbf{M}_{\text{FMN}} + \frac{1}{n}\sum_{i=1}^{n}\big(\mathbf{J}_{\text{FMN}}^{(1)}(\mathbf{x}_i, y_i) - 2\mathbf{M}_{\text{FMN}}\big) + o_p\big(n^{-1/2}\big),$$

where

$$\begin{aligned}\mathbf{J}_{\text{FMN}}^{(1)}(\mathbf{x}, y) = E_{(\mathbf{U}_2,V_2)}\big[&\mathbf{J}_{\text{FMN}}((\mathbf{x}, y), (\mathbf{U}_2, V_2))\\ &+ \mathbf{J}_{\text{FMN}}((\mathbf{x}, y), (\mathbf{U}_2, V_2))^\tau\big].\end{aligned}$$

Let $\mathbf{\Sigma}_{\text{FMN}}$ be the covariance matrix of $\text{vec}(\mathbf{J}_{\text{FMN}}^{(1)}(\mathbf{X}, Y))$. Then

$$\sqrt{n}\big(\text{vec}(\hat{\mathbf{M}}_{\text{FMN}}) - \text{vec}(\mathbf{M}_{\text{FMN}})\big) \xrightarrow{\mathcal{L}} \text{N}(\mathbf{0}, \mathbf{\Sigma}_{\text{FMN}}), \quad \text{as } n \to \infty.$$

5. IMPLEMENTATION

In this section we describe the procedures for estimating $\mathcal{S}_{Y|\mathbf{X}}$ and $\mathcal{S}_{E[Y|\mathbf{X}]}$ using the estimated candidate matrices $\hat{\mathbf{M}}_{\text{FCN}}$ and $\hat{\mathbf{M}}_{\text{FMN}}$. We also discuss the determination of dimensionality q and the choice of tuning parameters σ_{W}^2 and σ_{T}^2.

5.1 Algorithms

If the dimensionality of $\mathcal{S}_{Y|\mathbf{X}}$ (or $\mathcal{S}_{E[Y|\mathbf{X}]}$) is known to be q, then the first q eigenvectors of $\mathbf{M}_{\text{FCN}}$ (or $\mathbf{M}_{\text{FMN}}$) form an orthogonal basis of $\mathcal{S}_{Y|\mathbf{X}}$ (or $\mathcal{S}_{E[Y|\mathbf{X}]}$). From the preceding sections, it is known that the eigenvectors of $\hat{\mathbf{M}}_{\text{FCN}}$ (or $\hat{\mathbf{M}}_{\text{FMN}}$) converge to those of $\mathbf{M}_{\text{FCN}}$ (or $\mathbf{M}_{\text{FMN}}$) at the rate of $\sqrt{n}$. Therefore, we can use the first q eigenvectors of $\hat{\mathbf{M}}_{\text{FCN}}$ (or $\hat{\mathbf{M}}_{\text{FMN}}$) to generate a linear subspace and use it as an estimate of $\mathcal{S}_{Y|\mathbf{X}}$ (or $\mathcal{S}_{E[Y|\mathbf{X}]}$), which is denoted by $\hat{\mathcal{S}}_{Y|\mathbf{X}}$ (or $\hat{\mathcal{S}}_{E[Y|\mathbf{X}]}$). In practice, $\mathbf{X}$ does not necessarily have mean $\mathbf{0}$ and identity covariance matrix, so the data first need to be standardized. Standardization generally does not affect the convergence rate of the estimates, but may change their asymptotic variances. The procedure for deriving $\hat{\mathcal{S}}_{Y|\mathbf{X}}$ (or $\hat{\mathcal{S}}_{E[Y|\mathbf{X}]}$) is summarized as follows:

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

The foregoing procedure is fairly standard in sufficient dimension reduction, except that the estimated candidate matrices $\hat{\mathbf{M}}_{\text{FCN}}$ (or $\hat{\mathbf{M}}_{\text{FMN}}$) based on the Fourier method is used in step 2. For convenience, in the rest of the article we refer to the procedure with $\hat{\mathbf{M}}_{\text{FCN}}$ simply as FC, representing the Fourier method for estimating the central subspace, and the procedure with $\hat{\mathbf{M}}_{\text{FMN}}$ as FM, representing the Fourier method for estimating the central mean subspace. Two parameters, q and σ_{W}^2, need to be specified in FM, whereas in FC, an additional parameter σ_{T}^2 must be chosen. The parameter q is different from the other two in that the former is a model parameter and the latter are tuning parameters. The foregoing procedures 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 $\mathcal{S}_{Y|\mathbf{X}}$ (or $\mathcal{S}_{E[Y|\mathbf{X}]}$) is unknown and must be inferred from data. One informal method is to generate the scree plot of the eigenvalues of $\hat{\mathbf{M}}_{\text{FCN}}$ (or $\hat{\mathbf{M}}_{\text{FMN}}$) as in principal components 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 have been proposed in the literature. For example, Li (1991, 1992) proposed using a chi-squared statistic to sequentially test $q = 0, 1, 2,$ and so on, whereas Cook and Yin (2001) advocated using permutation tests for the same purpose. Recently, Ye and Weiss (2003) proposed using the bootstrap procedure to determine q. Next, we follow the basic idea of Ye and Weiss (2003) to develop the bootstrap procedure for choosing q in FC (or FM).

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

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

1. Randomly resample from $\{(\mathbf{x}_i, y_i)\}_{1\le i\le n}$ with replacement to generate N bootstrap samples each of size n, and denote the jth sample by $\{(\mathbf{x}_i^{(j)}, y_i^{(j)})\}_{1\le i\le n}$ for $1 \le j \le N$.
2. Based on each bootstrap sample [e.g., the jth sample $\{(\mathbf{x}_i^{(j)}, y_i^{(j)})\}_{1\le i\le n}$], 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 this procedure for $q = 1, \dots, p$ results in $\{\bar{d}(q)\}_{q=1}^p$, which are the variabilities of $\hat{\mathcal{S}}_q$ for $1 \le q \le p$.

Suppose that 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 larger 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 growing larger 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 0. When $q = q_0$, $\hat{\mathcal{S}}_q$ and $\hat{\mathcal{S}}_q^{(j)}$ estimate the same fixed space $\mathcal{S}$, and 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 $\bar{d}(q)$, and then decreases to 0 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 are 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 the bandwidths of the weight functions $K(\boldsymbol{\omega})$ and $k(t)$. However, the selection of σ_{W} and σ_{T} is fundamentally different from the selection of bandwidths for kernels used in nonparametric function estimation. In the latter case, the bandwidth of a kernel needs to decrease to 0 to ensure the asymptotic consistency of the estimated function as the sample size goes to infinity. In FC and FM, the consistency of $\hat{\mathbf{M}}_{\mathrm{FMN}}$ and $\hat{\mathbf{M}}_{\mathrm{FCN}}$ hold for any fixed positive σ_{W}^2 and σ_{T}^2; see Propositions 3 and 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 what follows, we first discuss the heuristics for choosing σ_{W}^2 and σ_{T}^2 and give their recommended values, then briefly introduce a bootstrap procedure for their optimal selection again following the idea of Ye and Weiss (2003).

We first discuss σ_{W}^2. When σ_{W}^2 is too large, $\boldsymbol{\phi}(\boldsymbol{\omega}, t)$ with large $\|\boldsymbol{\omega}\|$ will receive much larger weight than when σ_{W}^2 is small. As explained earlier, $\boldsymbol{\phi}(\boldsymbol{\omega}, t)$ with large $\|\boldsymbol{\omega}\|$ 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 (e.g., close to 0), then the weight assigned to $\boldsymbol{\phi}(\boldsymbol{\omega}, t)$ is almost 0 except for $\boldsymbol{\omega}$ 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, $\sigma_{\mathrm{W}}^2 = .1$) generally works well for standardized data. Similarly, for FM, we also recommend using $\sigma_{\mathrm{W}}^2 = .1$ in calculating $\hat{\mathbf{M}}_{\mathrm{FMN}}$.

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

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

In practice, some applications may require optimally choosing q, σ_{W}^2, and σ_{T}^2. We recommend the following procedure to use the bootstrap repeatedly. First, q is determined with $\sigma_{\mathrm{W}}^2 = .1$ and $\sigma_{\mathrm{T}}^2 = 1.0$ and denoted by q_1; second, σ_{W}^2 is chosen with $\sigma_{\mathrm{T}}^2 = 1.0$ and q_1 and denoted by σ_{W1}^2; third, σ_{T}^2 is selected with σ_{W1}^2 and q_1 and denoted by σ_{T1}^2. The steps are iterated until the parameters are stabilized. Our experience suggests that one iteration is usually sufficient. The second and third steps can be combined if a two-dimensional grid is used for selecting σ_{W}^2 and σ_{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 $\mathbf{X} = (X_1, \dots, X_{10})^\tau$ 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), $\boldsymbol{\beta}_1 = (1, 1, 1, 1, 0, 0, 0, 0, 0, 0)^\tau$ and $\boldsymbol{\beta}_2 = (0, 0, 0, 0, 0, 0, 1, 1, 1, 1)^\tau$. In the last example, we apply FC to a real dataset called 1985 Automobile Data to study how the price of car depends on its features.

Example 1. Consider the model $Y = (\boldsymbol{\beta}_1^\tau \mathbf{X})^2/(3 + (\boldsymbol{\beta}_2^\tau \mathbf{X} + 2)^2) + .2\varepsilon$. In this model the central subspace and the central mean subspace are identical, that is, $\mathcal{S}_{Y|\mathbf{X}} = \mathcal{S}_{E[Y|\mathbf{X}]} = \mathcal{S}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$. Let $\mathcal{S} = \mathcal{S}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$; thus, 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 ($\sigma_{\mathrm{W}}^2 = .1$, $\sigma_{\mathrm{T}}^2 = 1.0$) and FM ($\sigma_{\mathrm{W}}^2 = .1$) 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 earlier 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 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 $E[Y|\mathbf{X}]$ and $\mathbf{X}$.

Example 2. Consider the heteroscedastic model $Y = (\boldsymbol{\beta}_1^\tau \mathbf{X}) + 4(\boldsymbol{\beta}_2^\tau \mathbf{X})\varepsilon$. For this model, the central mean subspace and the central subspace are different, because $\mathcal{S}_{E[Y|\mathbf{X}]} = \mathcal{S}(\boldsymbol{\beta}_1)$ and $\mathcal{S}_{Y|\mathbf{X}} = \mathcal{S}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$. Clearly, $\mathcal{S}_{E[Y|\mathbf{X}]}$ is only a proper subspace of $\mathcal{S}_{Y|\mathbf{X}}$, so we focus our attention on $\mathcal{S}_{Y|\mathbf{X}}$ and the methods aimed at estimating $\mathcal{S}_{Y|\mathbf{X}}$. We draw a sample of 500 data points and apply FC ($\sigma_{\mathrm{W}}^2 = .1$ and $\sigma_{\mathrm{T}}^2 = 1.0$) to estimate $\mathcal{S}_{Y|\mathbf{X}}$. Only the first two eigenvalues of the estimated candidate matrix $\hat{\mathbf{M}}_{\mathrm{FCN}}$ are relatively large. We use the bootstrap procedure to generate the dimension variability plot based on 500 bootstrap samples, which is shown in Figure 2(a). Using the rule described in Section 5.2, it confirms that the dimension of $\mathcal{S}_{Y|\mathbf{X}}$ is equal to 2. Therefore, $\mathcal{S}_{Y|\mathbf{X}}$ is estimated by the space spanned the first two eigenvectors of $\hat{\mathbf{M}}_{\mathrm{FCN}}$, which are

$$\hat{\boldsymbol{\beta}}_1^\tau = (.5012, .4414, .4869, .4015, -.0205, .2531, -.1111, .0366, -.0709, .0266)$$

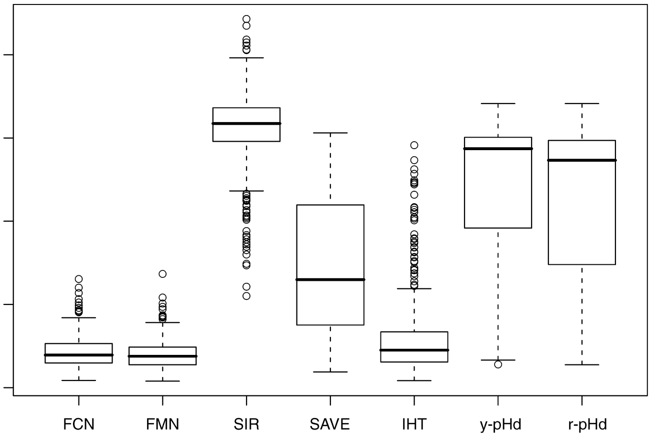

Figure 1. Side-by-Side Boxplots for the Performance Comparison Between FC, FM, SIR, SAVE, IHT, y-pHd, and r-pHd in Example 1. The y-axis represents the distance between an estimated subspace and the true subspace. Each boxplot is based on 500 samples.

and

$$\hat{\boldsymbol{\beta}}_2^\tau = (.0381, .2085, -.0590, .0025, -.1442, -.0137, .4853, .5143, .5186, .3658).$$

The distance between $\hat{\mathcal{S}}_{Y|\mathbf{X}} = \mathcal{S}(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$ and the true $\mathcal{S}_{Y|\mathbf{X}}$ is .04388.

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 $\mathcal{S}_{Y|\mathbf{X}}$. The distances are summarized by the side-by-side boxplots included in Figure 2(b). The boxplots indicate that FC outperforms both SIR and SAVE in this example.

Example 3. Consider the following model with discrete response: $Y = I_{[\boldsymbol{\beta}_1^\tau \mathbf{X} + \sigma\varepsilon > 1]} + 2I_{[\boldsymbol{\beta}_2^\tau \mathbf{X} + \sigma\varepsilon > 0]}$, where $I_{[\cdot]}$ denotes the indicator function and $\sigma = .2$. So the possible values for Y are 0, 1, 2, and 3. In this example we consider only the central subspace, which is spanned by $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. It is clear that Y is not a continuous function of $\mathbf{X}$. In the derivation of FC, a few differentiability conditions are required. But in the final formula for $\mathbf{M}_{\mathrm{FCN}}$, no differentiation is involved. So we expect FC would still work in this example. In fact, the involved differentiability in deriving FC is required only in a generalized sense. We draw a sample of 500 points and apply FC ($\sigma_{\mathrm{W}}^2 = .1$ and $\sigma_{\mathrm{T}}^2 = 3.0$) to estimate $\mathcal{S}_{Y|\mathbf{X}}$. Here 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 $\hat{\mathbf{M}}_{\mathrm{FCN}}$ are relatively large, indicating that the dimension of $\mathcal{S}_{Y|\mathbf{X}}$ is 2. This is further confirmed by the dimension variability plot included in Figure 3(a). Therefore, the estimated central subspace is spanned by the first two eigenvectors of $\hat{\mathbf{M}}_{\mathrm{FCN}}$, which are

$$\hat{\boldsymbol{\beta}}_1^\tau = (.0603, .0927, .0567, .0521, -.0432, -.0106, .4964, .3776, .4923, .4571)$$

and

$$\hat{\boldsymbol{\beta}}_2^\tau = (.5274, .5209, .5306, .4528, .0580, .1041, -.0735, -.0771, -.0729, -.0819).$$

The distance between $\mathcal{S}(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$ and the true subspace $\mathcal{S}_{Y|\mathbf{X}}$ is $d = .007809$.

We use the same procedure as in Example 2 to compare FC ($\sigma_{\mathrm{W}}^2 = .1$ 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 shown in Figure 3(b). 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 considered an extension of linear discriminant analysis.

Example 4. In this example we use FC to analyze a real dataset known as 1985 Automobile Data. The objective is to study how the price of car depends on its features. The dataset is available at the UCI Machine Learning Repository (*ftp://ftp.ics.uci.edu/pub/machine-learning-databases/autos*). Originally,

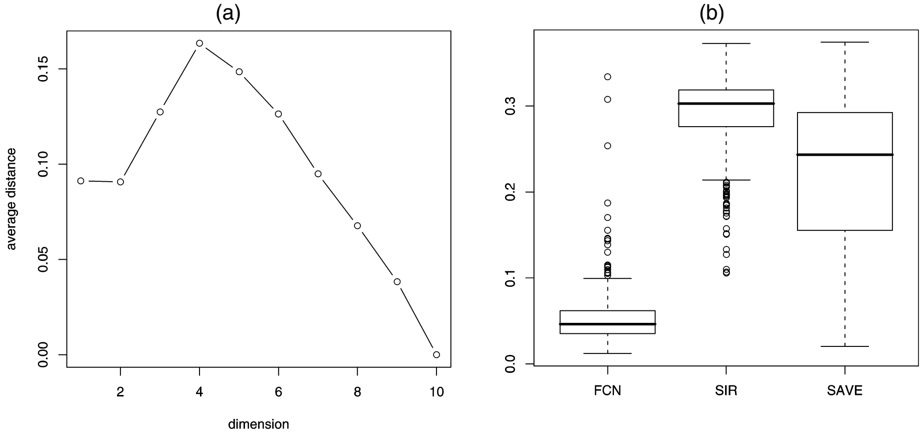

Figure 2. Dimension Variability Plot of $\bar{d}(q)$ versus q Based on 500 Bootstrap Samples in Example 2 (a) and Side-by-Side Boxplots for the Performance Comparison Between FC, SIR, and SAVE in Example 2 (b). The y-axis represents the distance between an estimated subspace and the true subspace. Each boxplot is based on 500 samples.

there were 205 instances (or cases) and 26 attributes (or variables) in the dataset, and there were some missing values. Because most current dimension-reduction methods, including FC, can only handle continuous variables, we remove eight categorical variables from the dataset. We remove one continuous variable that contains many missing values. For simplicity, we also discard the instances with missing values. The resulting dataset contains 195 instances and 14 variables: 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 predictor using its mean and standard deviation.

We first use FC with $\sigma_{\mathrm{W}}^2 = .1$ 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. To obtain sharper views, we fix $q = 2$ and use the bootstrap procedure mentioned at 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. Using FC with the tuned parameters, we calculate $\hat{\mathbf{M}}_{\mathrm{FCN}}$ and derive its spectral decomposition. The first two eigenvalues of $\hat{\mathbf{M}}_{\mathrm{FCN}}$, $\hat{\lambda}_1 = .1071$ and $\hat{\lambda}_2 = .0381$, are dominant. The bootstrap procedure is run with 1,000 bootstrap samples, and the dimension variability plot is generated and shown in Figure 5(a). The plot suggests that $q = 2$, so we use the first two eigenvectors of $\hat{\mathbf{M}}_{\mathrm{FCN}}$ to derive the estimate $\hat{\mathcal{S}}_{Y|\mathbf{X}} = \mathcal{S}(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$, where

$$\hat{\boldsymbol{\beta}}_1^{\tau} = (.05, -.19, .08, .09, .75, -.24, \\ 0, -.11, .17, .51, .04, -.08, .12)$$

and

$$\hat{\boldsymbol{\beta}}_2^{\tau} = (.08, -.38, .09, .08, .03, .70, \\ -.17, -.20, -.06, -.08, .06, .49, -.17).$$

Figure 4 includes the projection plots of $\log(y)$ versus $\hat{\boldsymbol{\beta}}_1^{\tau}\mathbf{x}$ (a) and $\log(y)$ versus $\hat{\boldsymbol{\beta}}_2^{\tau}\mathbf{x}$ (b), with 4(a) displaying a strong linear relationship and 4(b) displaying a parabolic relationship.

Based on the relative magnitudes of the components, $\hat{\boldsymbol{\beta}}_1$ is determined mainly by Curb weight (x_5) and Horsepower (x_{10}),

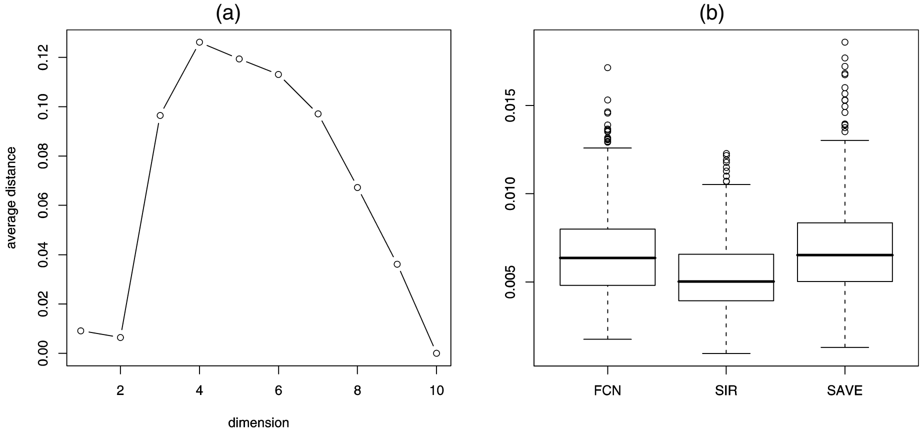

Figure 3. Dimension Variability Plot Based on 500 Bootstrap Samples in Example 3 (a) and Side-by-Side Boxplots for the Performance Comparison Between FC, SIR, and SAVE in Example 3 (b). The y-axis represents the distance between an estimated subspace and the true subspace. Each boxplot is based on 500 samples.

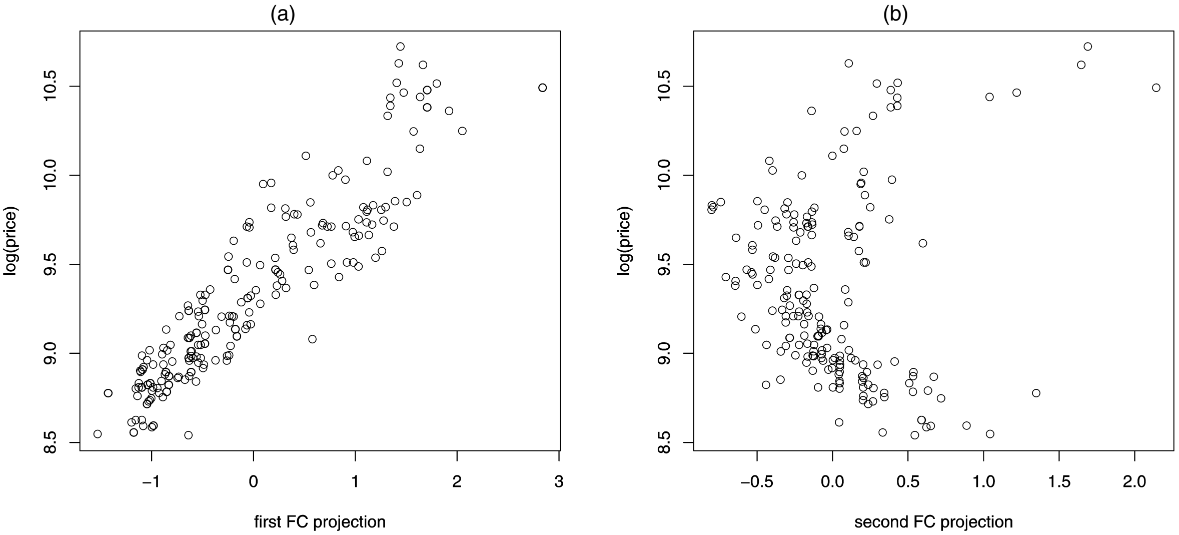

Figure 4. Plot of log(y) versus $\hat{\boldsymbol{\beta}}_1^\tau \mathbf{x}$ (a) and Plot of log(y) versus $\hat{\boldsymbol{\beta}}_2^\tau \mathbf{x}$ (b) in Example 4.

followed by the other variables, and $\hat{\boldsymbol{\beta}}_2$ is determined mainly by Engine size (x_6) and City mpg (x_{12}). The strong linear relationship between log(y) and $\hat{\boldsymbol{\beta}}_1^\tau \mathbf{x}$ indicates that the price of a car can be well predicted by Curb weight and Horsepower. The parabolic relationship between log(y) and $\hat{\boldsymbol{\beta}}_2^\tau \mathbf{x}$ reveals that the price of a car also depends on Engine size and City mpg, but 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, say, Mercedes-Benz, BMW, Jaguar, and Porsche, whereas the points in the lower branch represent lower-end cars, such as the hatchbacks of Honda, Chevrolet, Plymouth, Subaru, and so on. For the high-end cars, the price increases as $\hat{\boldsymbol{\beta}}_2^\tau \mathbf{x}$ increases, whereas for the lower-end cars, the price decreases as $\hat{\boldsymbol{\beta}}_2^\tau \mathbf{x}$ increases. The two relationships above further imply that there exists a nonlinear confounding between $\hat{\boldsymbol{\beta}}_1^\tau \mathbf{x}$ and $\hat{\boldsymbol{\beta}}_2^\tau \mathbf{x}$. Figure 5(b) plots $\hat{\boldsymbol{\beta}}_1^\tau \mathbf{x}$ versus $\hat{\boldsymbol{\beta}}_2^\tau \mathbf{x}$, which exhibits a parabolic relationship between these two directions. (See 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, whereas the latter fails to recover interesting directions in this particular example.

7. CONCLUSION

Using the Fourier transform, we have derived two candidate matrices, $\mathbf{M}_{\mathrm{FC}}$ and $\mathbf{M}_{\mathrm{FM}}$, for recovering the entire central subspace and central mean subspace. Under further distributional assumptions, we have derived explicit estimates of the candidate matrices that lead to estimates of the central and central mean subspaces. Selection of the tuning parameters and determination of dimensionality have been discussed. Synthetic and real examples were used to demonstrate the performance of the proposed methods in comparison with other methods. Use of

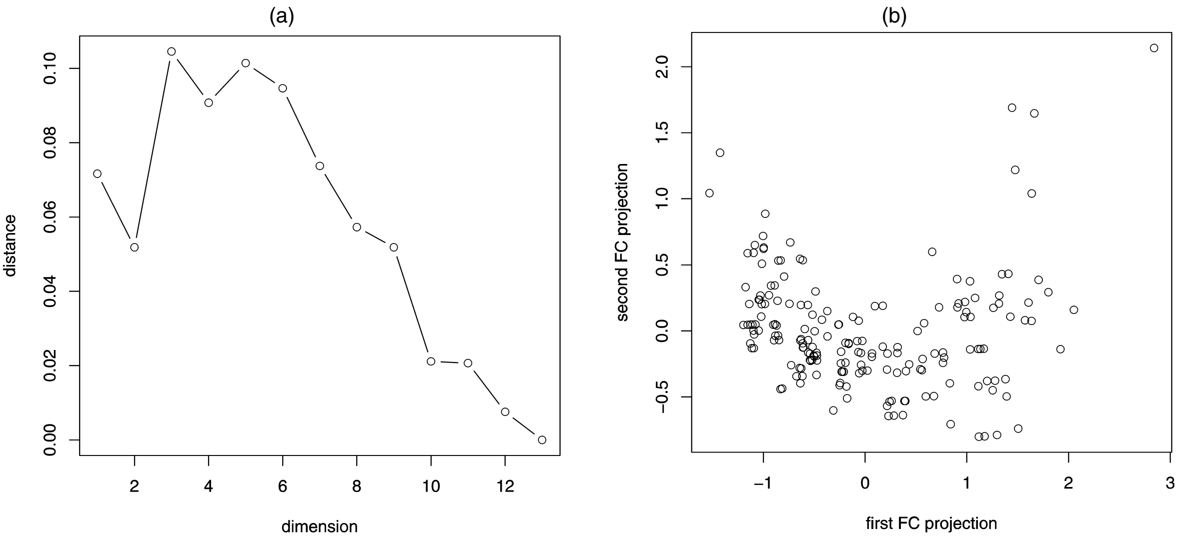

Figure 5. The Dimension Variability Plot Generated From the Bootstrap Procedure With 1,000 Bootstrap Samples, $\sigma_W^2 = .08$ and $\sigma_T^2 = .6$ (a) and Plot of $\hat{\boldsymbol{\beta}}_1^\tau \mathbf{x}$ versus $\hat{\boldsymbol{\beta}}_2^\tau \mathbf{x}$ That Shows a Nonlinear Relationship (b). Both plots are for Example 4.

the Fourier transform may provide a different approach to dimension reduction in general regression, which is expected to generate more interesting results in the future. Currently we are focused on two issues: (1) to generalize the results of this article to the case without distributional assumptions imposed on $\mathbf{X}$ and (2) 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 than SIR and SAVE for the latter case.

APPENDIX: PROOFS

Proof of Equation (5)

Because $\mathbf{A} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \dots, \boldsymbol{\alpha}_q)$ and the columns $\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_q$ form a basis for $\mathcal{S}_{E[Y|\mathbf{X}]}$ with q dimensions, to prove (5), it is sufficient to show that for $\boldsymbol{\beta} \in \mathbb{R}^p$, $\boldsymbol{\beta}^\tau \mathbf{A} = \mathbf{0}$ is equivalent to $\boldsymbol{\beta}^\tau \frac{\partial}{\partial \mathbf{x}} m(\mathbf{x}) = 0$ for all $\mathbf{x} \in \operatorname{supp}(\mathbf{X})$.

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

Next, we show that the converse is also true by contradiction. Assume that there exists $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ such that $\boldsymbol{\beta}_0^\tau \frac{\partial}{\partial \mathbf{x}} m(\mathbf{x}) = 0$ for all $\mathbf{x} \in \operatorname{supp}(\mathbf{X})$, but $\boldsymbol{\beta}_0^\tau \mathbf{A} \neq \mathbf{0}$. Let $\boldsymbol{\xi}_1 = \mathbf{A}^\tau \boldsymbol{\beta}_0 / \|\mathbf{A}^\tau \boldsymbol{\beta}_0\|$. Clearly, $\boldsymbol{\xi}_1$ is a q-dimensional nonzero vector. So $\boldsymbol{\beta}_0^\tau \frac{\partial}{\partial \mathbf{x}} m(\mathbf{x}) = \boldsymbol{\beta}_0^\tau \mathbf{A} \frac{\partial}{\partial \mathbf{u}} g(\mathbf{u}) = 0$ means that $\boldsymbol{\xi}_1^\tau \frac{\partial}{\partial \mathbf{u}} g(\mathbf{u}) = 0$, which implies that the directional derivative of g as a function of $\mathbf{u} = (u_1, \dots, u_q)^\tau$ along $\boldsymbol{\xi}_1$ is always 0. This further implies that $g(\mathbf{u})$ is a constant along $\boldsymbol{\xi}_1$, that is, $g(\mathbf{u} + t\boldsymbol{\xi}_1) = g(\mathbf{u})$ for $t \in \mathbb{R}$. We can expand $\boldsymbol{\xi}_1$ by bringing in $\boldsymbol{\xi}_2, \dots, \boldsymbol{\xi}_q$ to form an orthonormal basis for $\mathbb{R}^q$. Let $\mathbf{D} = (\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_q)$ and $\mathbf{v} = \mathbf{D}^\tau \mathbf{u} = (v_1, \dots, v_q)^\tau$; then $g(\mathbf{u}) = g(\mathbf{D}\mathbf{v})$ and $\frac{\partial}{\partial v_1} g(\mathbf{D}\mathbf{v}) = \boldsymbol{\xi}_1^\tau \frac{\partial}{\partial \mathbf{u}} g(\mathbf{u}) = 0$. Hence g does not depend on v_1, so we can rewrite $g(\mathbf{u}) = g(\mathbf{D}\mathbf{v}) = \tilde{g}(v_2, \dots, v_q) = \tilde{g}(\boldsymbol{\xi}_2^\tau \mathbf{A}^\tau \mathbf{x}, \dots, \boldsymbol{\xi}_q^\tau \mathbf{A}^\tau \mathbf{x})$, which implies that $\mathcal{S}(\mathbf{A}\boldsymbol{\xi}_2, \dots, \mathbf{A}\boldsymbol{\xi}_q)$ is also a dimension-reduction subspace for $E[Y|\mathbf{X}]$ and that the central mean subspace has dimension at most $q-1$, which contradicts to $\dim(\mathcal{S}_{E[Y|\mathbf{X}]}) = q$. Thus the proof is completed.

Proof of Proposition 1

1. From the definition of $\boldsymbol{\psi}(\boldsymbol{\omega})$ in (6), $\mathbf{a}(\boldsymbol{\omega})$ and $\mathbf{b}(\boldsymbol{\omega})$ have the following expressions:

$$\mathbf{a}(\boldsymbol{\omega}) = \int \cos(\boldsymbol{\omega}^\tau \mathbf{x}) \left(\frac{\partial}{\partial \mathbf{x}} m(\mathbf{x}) \right) f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}$$

and

$$\mathbf{b}(\boldsymbol{\omega}) = \int \sin(\boldsymbol{\omega}^\tau \mathbf{x}) \left(\frac{\partial}{\partial \mathbf{x}} m(\mathbf{x}) \right) f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}.$$

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

2. Because $m(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) \to 0$ as $\|\mathbf{x}\| \to \infty$, using integration by parts, we have

$$\begin{aligned}\boldsymbol{\psi}(\boldsymbol{\omega}) &= -\int m(\mathbf{x}) \Big(\imath \boldsymbol{\omega} \exp\{\imath \boldsymbol{\omega}^\tau \mathbf{x}\} f_{\mathbf{X}}(\mathbf{x}) \\ &\qquad + \exp\{\imath \boldsymbol{\omega}^\tau \mathbf{x}\} \frac{\partial}{\partial \mathbf{x}} f_{\mathbf{X}}(\mathbf{x}) \Big)\, d\mathbf{x} \\ &= -E_{(\mathbf{X},Y)}\big[(\imath \boldsymbol{\omega} + \mathbf{G}(\mathbf{X})) Y \exp\{\imath \boldsymbol{\omega}^\tau \mathbf{X}\} \big].\end{aligned}$$

3. The proof is standard and has been given by Folland (1992, pp. 217, 243).
4. The proof is standard and has been given by Folland (1992, pp. 222, 244).

Proof of Proposition 2

Because $\boldsymbol{\psi}(\boldsymbol{\omega}) = \mathbf{a}(\boldsymbol{\omega}) + \imath \mathbf{b}(\boldsymbol{\omega})$, we have

$$\begin{aligned}\boldsymbol{\psi}(\boldsymbol{\omega}) \bar{\boldsymbol{\psi}}(\boldsymbol{\omega})^\tau &= [\mathbf{a}(\boldsymbol{\omega})\mathbf{a}(\boldsymbol{\omega})^\tau + \mathbf{b}(\boldsymbol{\omega})\mathbf{b}(\boldsymbol{\omega})^\tau] \\ &\qquad + \imath [\mathbf{b}(\boldsymbol{\omega})\mathbf{a}(\boldsymbol{\omega})^\tau - \mathbf{a}(\boldsymbol{\omega})\mathbf{b}(\boldsymbol{\omega})^\tau].\end{aligned}$$

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

$$\begin{aligned}\mathbf{M}^*_{\mathrm{FM}} &= (2\pi)^{-p} \int \boldsymbol{\psi}(\boldsymbol{\omega}) \bar{\boldsymbol{\psi}}(\boldsymbol{\omega})^\tau\, d\boldsymbol{\omega} \\ &= (2\pi)^{-p} \int [\mathbf{a}(\boldsymbol{\omega})\mathbf{a}(\boldsymbol{\omega})^\tau + \mathbf{b}(\boldsymbol{\omega})\mathbf{b}(\boldsymbol{\omega})^\tau]\, d\boldsymbol{\omega}.\end{aligned}$$

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

Proof of Proposition 3

Because

$$\mathbf{M}_{\mathrm{FM}} = \int [\mathbf{a}(\boldsymbol{\omega})\mathbf{a}(\boldsymbol{\omega})^\tau + \mathbf{b}(\boldsymbol{\omega})\mathbf{b}(\boldsymbol{\omega})^\tau] K(\boldsymbol{\omega})\, d\boldsymbol{\omega},$$

$\mathbf{M}_{\mathrm{FM}}$ is a nonnegative definite matrix. Because $K(\boldsymbol{\omega}) > 0$ for $\boldsymbol{\omega} \in \mathbb{R}^p$, for $\boldsymbol{\beta} \in \mathbb{R}^p$ we have

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

where $\Longleftrightarrow$ means "being equivalent to." Therefore, $\mathcal{S}(\mathbf{M}_{\mathrm{FM}}) = \operatorname{span}\{\mathbf{a}(\boldsymbol{\omega}), \mathbf{b}(\boldsymbol{\omega}) : \boldsymbol{\omega} \in \mathbb{R}^p\} = \mathcal{S}_{E[Y|\mathbf{X}]}$.

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

$$\mathcal{S}_{Y|\mathbf{X}} = \operatorname{span}\left\{ \frac{\partial}{\partial \mathbf{x}} f_{Y|\mathbf{X}}(y|\mathbf{x}) : (\mathbf{x}, y) \in \operatorname{supp}(\mathbf{X}, Y) \right\},$$

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

$$\begin{aligned}&\boldsymbol{\beta}^\tau \frac{\partial}{\partial \mathbf{x}} m(\mathbf{x}, t) = 0 \quad \text{for all } t \in \mathbb{R} \\ &\qquad \Longleftrightarrow \quad \boldsymbol{\beta}^\tau \frac{\partial}{\partial \mathbf{x}} f_{Y|\mathbf{X}}(y|\mathbf{x}) = 0 \quad \text{for all } y \in \operatorname{supp}(Y).\end{aligned}$$

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

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

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 $m(\mathbf{x})$ replaced by $m(\mathbf{x}, t)$. The details are thus omitted.

Proof of Proposition 6

Because

$$\mathbf{M}_{\mathrm{FC}} = \iint [\mathbf{a}(\boldsymbol{\omega}, t)\mathbf{a}(\boldsymbol{\omega}, t)^{\tau} + \mathbf{b}(\boldsymbol{\omega}, t)\mathbf{b}(\boldsymbol{\omega}, t)^{\tau}]K(\boldsymbol{\omega})k(t)\, d\boldsymbol{\omega}\, dt,$$

$\mathbf{M}_{\mathrm{FC}}$ is a nonnegative definite matrix. Because $K(\boldsymbol{\omega}) > 0$ for $\boldsymbol{\omega} \in \mathbb{R}^p$ and $k(t) > 0$ for $t \in \mathbb{R}$, we have, for any $\boldsymbol{\beta} \in \mathbb{R}^p$,

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

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

Proof of Proposition 7

Applying integration by parts and the condition that $f_{(\mathbf{X},Y)}(\mathbf{x}, y)$ goes to 0 as $\mathbf{x}$ goes to infinity and y is fixed, we have

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

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

Proof of Proposition 8

When $\mathbf{X}$ follows the normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}_p$, $\mathbf{G}(\mathbf{X}) = -\mathbf{X}$. Because $K(\boldsymbol{\omega})$ is a point mass at $\boldsymbol{\omega} = \mathbf{0}$,

$$\mathbf{M}_{\mathrm{FC}} = \int [\mathbf{a}(\mathbf{0}, t)\mathbf{a}(\mathbf{0}, t)^{\tau} + \mathbf{b}(\mathbf{0}, t)\mathbf{b}(\mathbf{0}, t)^{\tau}]k(t)\, dt,$$

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

$$\begin{aligned}
\boldsymbol{\phi}(\mathbf{0}, t) &= \int \exp\{\imath ty\}\mathbf{x}f_{(\mathbf{X},Y)}(\mathbf{x}, y)\, d\mathbf{x}\, dy \\
&= \int \exp\{\imath ty\}E[\mathbf{X}|Y = y]f_Y(y)\, dy.
\end{aligned}$$

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

Proof of Theorem 1

Because $\hat{\mathbf{M}}_{\mathrm{FCN}}$ is a V-statistic, it can be written as

$$\begin{aligned}
\hat{\mathbf{M}}_{\mathrm{FCN}} &= \frac{n-1}{2n}\binom{n}{2}^{-1} \\
&\quad \times \sum_{i<j}\{\mathbf{J}_{\mathrm{FCN}}((\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j)) + \mathbf{J}_{\mathrm{FCN}}^{\tau}((\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j))\} \\
&\quad + \frac{1}{n^2}\sum_{i=1}^{n}\mathbf{J}_{\mathrm{FCN}}((\mathbf{x}_i, y_i), (\mathbf{x}_i, y_i)).
\end{aligned}$$

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

$$\begin{aligned}
\hat{\mathbf{M}}_{\mathrm{FCN}} &= \frac{n-1}{2n}\Biggl(2\mathbf{M}_{\mathrm{FCN}} + \frac{2}{n}\sum_{i=1}^{n}\bigl(\mathbf{J}_{\mathrm{FCN}}^{(1)}(\mathbf{x}_i, y_i) - 2\mathbf{M}_{\mathrm{FCN}}\bigr) \\
&\qquad + o_p\bigl(n^{-1/2}\bigr)\Biggr) + O_p(n^{-1}) \\
&= \mathbf{M}_{\mathrm{FCN}} + \frac{1}{n}\sum_{i=1}^{n}\bigl(\mathbf{J}_{\mathrm{FCN}}^{(1)}(\mathbf{x}_i, y_i) - 2\mathbf{M}_{\mathrm{FCN}}\bigr) + o_p\bigl(n^{-1/2}\bigr).
\end{aligned}$$

The second term in the foregoing expression is the average of n independent and identically distributed random matrices $(\mathbf{J}_{\mathrm{FCN}}^{(1)}(\mathbf{x}_i, y_i) - 2\mathbf{M}_{\mathrm{FCN}})$ with $1 \le i \le n$. Because $\boldsymbol{\Sigma}_{\mathrm{FCN}}$ exists, which is guaranteed by the existence of the covariance matrix of $\operatorname{vec}(\mathbf{J}_{\mathrm{FCN}}((\mathbf{U}_1, V_1), (\mathbf{U}_2, V_2)))$, by the central limit theorem, as n goes to infinity,

$$\sqrt{n}\bigl(\operatorname{vec}(\hat{\mathbf{M}}_{\mathrm{FCN}}) - \operatorname{vec}(\mathbf{M}_{\mathrm{FCN}})\bigr) \xrightarrow{\mathcal{L}} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathrm{FCN}}),$$

where $\boldsymbol{\Sigma}_{\mathrm{FCN}} = \operatorname{cov}[\operatorname{vec}(\mathbf{J}_{\mathrm{FCN}}^{(1)}(\mathbf{X}, Y))]$ is a $p^2 \times p^2$ matrix.

Proof of Proposition 9

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

$$\begin{aligned}
\boldsymbol{\eta}(y, \boldsymbol{\omega}) &= \int \exp\{\imath\boldsymbol{\omega}^{\tau}\mathbf{x}\}\frac{\partial}{\partial\mathbf{x}}f_{\mathbf{X}|Y}(\mathbf{x}|y)\, d\mathbf{x} + \int \mathbf{x}\exp\{\imath\boldsymbol{\omega}^{\tau}\mathbf{x}\}f_{\mathbf{X}|Y}(\mathbf{x}|y)\, d\mathbf{x} \\
&= f_Y(y)^{-1}\int \exp\{\imath\boldsymbol{\omega}^{\tau}\mathbf{x}\}\left(\frac{\partial}{\partial\mathbf{x}}f_{\mathbf{X}}(\mathbf{x}) + \mathbf{x}f_{\mathbf{X}}(\mathbf{x})\right)f_{Y|\mathbf{X}}(y|\mathbf{x})\, d\mathbf{x} \\
&\quad + f_Y(y)^{-1}\int \exp\{\imath\boldsymbol{\omega}^{\tau}\mathbf{x}\}\left(\frac{\partial}{\partial\mathbf{x}}f_{Y|\mathbf{X}}(y|\mathbf{x})\right)f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}.
\end{aligned}$$

The second term belongs to $\mathcal{S}_{Y|\mathbf{X}}$, so we need only focus on the first term and denote it by $\mathbf{I}_1$. Let $\mathbf{u}_1 = \mathbf{B}^{\tau}\mathbf{x}$, $\mathbf{u}_2 = \tilde{\mathbf{B}}^{\tau}\mathbf{x}$, and $\mathbf{u} = \binom{\mathbf{u}_1}{\mathbf{u}_2} = (\mathbf{B}, \tilde{\mathbf{B}})^{\tau}\mathbf{x} = \mathbf{M}^{\tau}\mathbf{x}$. Then

$$\begin{aligned}
\mathbf{I}_1 &= f_Y^{-1}(y)\int \exp\{\imath\tilde{\boldsymbol{\omega}}^{\tau}\mathbf{u}\}\mathbf{M}\left(\frac{\partial}{\partial\mathbf{u}}\tilde{p}(\mathbf{u}) + u\tilde{p}(\mathbf{u})\right)p(y|\mathbf{u}_1)\, d\mathbf{u} \\
&= f_Y^{-1}(y)\mathbf{B}\int \exp\{\imath\tilde{\boldsymbol{\omega}}^{\tau}\mathbf{u}\}\left(\frac{\partial}{\partial\mathbf{u}_1}\tilde{p}(\mathbf{u}) + \mathbf{u}_1\tilde{p}(\mathbf{u})\right)p(y|\mathbf{u}_1)\, d\mathbf{u} \\
&\quad + f_Y^{-1}(y)\tilde{\mathbf{B}}\int \exp\{\imath\tilde{\boldsymbol{\omega}}^{\tau}\mathbf{u}\}\left(\frac{\partial}{\partial\mathbf{u}_2}\tilde{p}(\mathbf{u}) + u_2\tilde{p}(\mathbf{u})\right) \\
&\quad \times p(y|\mathbf{u}_1)\, d\mathbf{u}_1\, d\mathbf{u}_2,
\end{aligned}$$

where $\tilde{\boldsymbol{\omega}} = \mathbf{M}^{\tau}\boldsymbol{\omega}$ and $\tilde{p}(\mathbf{u}) = f_{\mathbf{X}}(\mathbf{M}\mathbf{u})$ is the density function of $\mathbf{u}$. Note that the first term falls in $\mathcal{S}_{Y|\mathbf{X}}$. Under the weak normality condition, we have $\frac{\partial}{\partial\mathbf{u}_2}\tilde{p}(\mathbf{u}_2|\mathbf{u}_1) + \mathbf{u}_2\tilde{p}(\mathbf{u}_2|\mathbf{u}_1) = \mathbf{0}$, so the second term in the foregoing expression is $\mathbf{0}$. Therefore $\mathbf{I}_1 \in \mathcal{S}_{Y|\mathbf{X}}$. This proves the proposition.

Proof of Theorem 2

The proof is similar to that of Theorem 1 and thus is omitted.

[Received July 2004. Revised November 2005.]

REFERENCES

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

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

——— (1998), *Regression Graphics: Ideas for Studying Regressions Through Graphics*, New York: Wiley.

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*, Belmont: Brooks/Cole.

Friedman, J. H., and Stuetzle, W. (1981), "Projection Pursuit Regression," *Journal of the American Statistical Association*, 76, 817–823.

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.

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

Hotelling, H. (1957), "The Relations of the Newer Multivariate Statistical Methods to Factor Analysis," *British Journal of Statistical Psychology*, 10, 69–79.

Kendall, M. G. (1957), *A Course in Multivariate Analysis*, London: Charles Griffin.

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

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

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

——— (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.

——— (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 discussion), *Journal of the Royal Statistical Society*, Ser. 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 *k*th Moment in Regression," *Journal of the Royal Statistical Society*, Ser. B, 64, 159–175.