Statistica Sinica **20** (2010), 1291-1307

ON DIMENSION REDUCTION IN REGRESSIONS WITH MULTIVARIATE RESPONSES

Li-Ping Zhu, Li-Xing Zhu and Song-Qiao Wen

East China Normal University, Hong Kong Baptist University and Shenzhen University

Abstract: This paper is concerned with dimension reduction in regressions with multivariate responses on high-dimensional predictors. A unified method that can be regarded as either an inverse regression approach or a forward regression method is proposed to recover the central dimension reduction subspace. By using Stein's Lemma, the forward regression estimates the first derivative of the conditional characteristic function of the response given the predictors; by using the Fourier method, the inverse regression estimates the subspace spanned by the conditional mean of the predictors given the responses. Both methods lead to an identical kernel matrix, while preserving as much regression information as possible. Illustrative examples of a data set and comprehensive simulations are used to demonstrate the application of our methods.

Key words and phrases: Central subspace, dimension reduction, ellipticity, inverse regression, multivariate response.

1. Introduction

Consider the regression of a multivariate response $\mathbf{Y} = (Y_1, \ldots, Y_q)^\tau$ on a p-dimensional predictor $\mathbf{X} = (X_1, \ldots, X_p)^\tau$, where the superscript "τ" denotes the transpose operator. When p is large, it is desirable to reduce the dimension of $\mathbf{X}$ to improve the efficacy of modeling. Toward this end, sufficient dimension reduction (SDR) is introduced to reduce the dimension of predictors without losing information of the regression of $\mathbf{Y}|\mathbf{X}$. The central subspace (CS, Cook (1998)), an important notion in the area of SDR, is defined as the smallest column space of $\mathbf{B}$ satisfying

$$\mathbf{Y} \perp\!\!\!\perp \mathbf{X} | \mathbf{B}^\tau \mathbf{X}. \tag{1.1}$$

Without notational confusion, we assume that the CS, denoted by $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$, is spanned by the columns of $\mathbf{B}$, namely, $\mathcal{S}_{\mathbf{Y}|\mathbf{X}} = \text{span}\{\mathbf{B}\}$. The dimension of CS, denoted by K in this context, is usually referred to as the structural dimension. Potential advantages accrue from working in the SDR context because it preserves the integrity of the regression information of $\mathbf{Y}$ given $\mathbf{X}$ without a pre-specified model. To reduce the dimension of predictors in the SDR context, there has been a focus on research in regressions with univariate response $\mathbf{Y}$. See, among many others, for example, sliced inverse regression (SIR, Li (1991)), slicing average variance estimation (SAVE, Cook and Weisberg (1991)), contour regression (SCR and GCR, Li, Zha, and Chiaromonte (2005)), directional regression (DR, Li and Wang (2007)), and references therein.

We consider general $q \geq 1$. For multivariate response data, the identification of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ has been developed in three main directions. The first estimates the joint CS directly by generalizing the slicing methodology of Li (1991) to multivariate $\mathbf{Y}$ (Aragon (1997); Hsing (1999); Setodji and Cook (2004)); similar to slicing the univariate response Y into intervals in the original form of SIR (Li (1991)), the multi-dimensional $\mathbf{Y}$ is divided into hypercubes. The joint slicing is effective when the dimension of $\mathbf{Y}$ is relatively small. However, as the dimension of $\mathbf{Y}$ increases, the number of observations within each hypercube decreases exponentially, which deteriorates the estimation efficacy. The second approach is to first estimate the marginal CS $\mathcal{S}_{Y_i|\mathbf{X}}$ for each coordinates of $\mathbf{Y}$, and then to combine all q marginal CS $\mathcal{S}_{Y_i|\mathbf{X}}$ to estimate the joint CS $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ (Cook and Setodji (2003); Saracco (2005); Yin and Bura (2006)). This does not guarantee recovery of the joint CS $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$. The third approach relies on classical methods; it is designed for a univariate response using one-dimensional projections of the original response $\mathbf{Y}$. Li et al. (2003) considered only a few projections of the responses, which may result in loss of information of the CS, and their method relies upon the choice of the initial value. We demonstrate this point with a simulated example in Section 5. Li, Wen, and Zhu (2008) proposed a random projection method with the benefit that multivariate slicing is avoided while the integrity of the joint CS $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ is preserved under some mild conditions. These projection methods, together with all previous ones, however, bring in the selection of tuning parameters, say, the number of slices in slicing estimation (Li (1991)), the bandwidth in kernel smoothing (Zhu and Fang (1996), Zhu and Zhu (2007)), or the number of knots in spline approximation (Zhu and Yu (2007)), etc. Because the performance of higher moments methods such as SAVE rely heavily on the choice of the number of slices (Cook and Critchley (2000), Zhu, Ohtaki, and Li (2007), Li and Zhu (2007)) even when the response is univariate, the tuning parameters must be selected delicately. However, the selection of an optimal tuning parameter is still an open problem.

The present paper represents another effort to reduce the dimension of predictors in regressions with multivariate responses. Our proposed approach is essentially a unification of inverse regression and forward regression that is called the UIF method. The forward regression estimates the first derivative of the conditional characteristic function of the response given the predictors, using Stein's Lemma, and the inverse regression estimates the subspace spanned by the conditional mean of the predictors given the responses, using Fourier methods. Both methods lead to an identical kernel matrix, while preserving as much regression information as possible.

We illustrate in detail the rationale of our UIF method by using the inverse regression $E(\mathbf{X}|\mathbf{Y})$, termed the UIF method using the first moment or, simply, UIF_1. The UIF_1 is on the basis of the seed vector

$$\boldsymbol{\phi}(\mathbf{t}) = E(e^{i\mathbf{t}^{\tau}\mathbf{Y}}\mathbf{X}) - E(e^{i\mathbf{t}^{\tau}\mathbf{Y}})E(\mathbf{X}) \tag{1.2}$$

whose real and imaginary parts, denoted by $\boldsymbol{\alpha}(\mathbf{t})$ and $\boldsymbol{\beta}(\mathbf{t})$ respectively, lie in $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ for any given $\mathbf{t} \in \mathcal{R}^q$ under very mild conditions. Clearly, the subspace $\mathrm{span}\{\boldsymbol{\alpha}(\mathbf{t}), \boldsymbol{\beta}(\mathbf{t}), \mathbf{t} \in \mathcal{R}^q\}$ is identical to the subspace $\mathrm{span}\{\mathbf{M}\}$, where

$$\mathbf{M} = \mathrm{real}\{E[\boldsymbol{\phi}(\mathbf{T})\bar{\boldsymbol{\phi}}^{\tau}(\mathbf{T})]\} = E[\boldsymbol{\alpha}(\mathbf{T})\boldsymbol{\alpha}^{\tau}(\mathbf{T}) + \boldsymbol{\beta}(\mathbf{T})\boldsymbol{\beta}^{\tau}(\mathbf{T})], \tag{1.3}$$

in which $\bar{\boldsymbol{\phi}}(\mathbf{T})$ is the conjugate of $\boldsymbol{\phi}(\mathbf{T})$, and $\mathbf{T}$ is a q-dimensional random vector whose support is $\mathcal{R}^q$. The notation $\mathrm{real}\{\boldsymbol{\gamma}\}$ stands for the real part of a complex matrix $\boldsymbol{\gamma}$. Consequently,

$$\mathrm{span}\{\mathbf{M}\} \subseteq \mathcal{S}_{\mathbf{Y}|\mathbf{X}}, \tag{1.4}$$

Theoretically, (1.4) holds for any $\mathbf{T}$ whose support is $\mathcal{R}^q$. In implementation, we design a data-adaptive mechanism to choose the random vector $\mathbf{T}$ to ensure estimation efficacy.

In recovering the CS it is also important to ask whether the candidate matrix $\mathbf{M}$ is capable of identifying CS exhaustively. It is known that SIR and UIF_1, using the first moment $E(\mathbf{X}|\mathbf{Y})$, fails to capture directions along which $\mathbf{Y}$ is symmetric. However SAVE, using the second moment $\mathrm{Cov}(\mathbf{X}|\mathbf{Y})$, can recover CS exhaustively when $\mathbf{X}|\mathbf{Y}$ is multivariate normal (Li and Wang (2007)). In this paper, we will explore our UIF method using the second moment, which we term the UIF_2 method.

The remainder of this paper is organized as follows. Section 2 motivates our UIF method and derives the candidate matrix $\mathbf{M}$ to identify the CS in the population level. Sections 3 turns to the implementation of our method for identifying the CS. Comprehensive simulation studies are reported in Section 4 to compare our method with others. Using the ideas of UIF_1, we also develop in Section 4 the UIF_2 method using the second conditional moment of $\mathbf{X}|\mathbf{Y}$. An analysis of a data set is reported in Section 5. The technical derivations are relegated to the Appendix.

The following notations are frequently used in our subsequent exposition. The notation "$\perp$" is the perpendicular operator in algebra, and "$\perp\!\!\!\perp$" denotes statistical independence. For a complex matrix $\boldsymbol{\gamma}$, the notation real$\{\boldsymbol{\gamma}\}$ stands for the real part of $\boldsymbol{\gamma}$, and $\overline{\boldsymbol{\gamma}}$ is the conjugate of $\boldsymbol{\gamma}$. Let $\mathbf{I}_p$ be the $p \times p$ identity matrix. The notation trace$(\mathbf{A})$ means the trace of a matrix $\mathbf{A}$, span$(\mathbf{A})$ denotes the subspace of $\mathcal{R}^p$ spanned by the columns of $\mathbf{A}$, and $P_{\mathbf{A}}$ is the projection operator in the standard inner product of $\mathbf{A}$, namely, $P_{\mathbf{A}} = \mathbf{A}(\mathbf{A}^\tau\mathbf{A})^{-1}\mathbf{A}^\tau$, and $\|\mathbf{A}\| = \sqrt{\text{trace}(\mathbf{A}^\tau\mathbf{A})}$.

2. Methodology Development

For ease of illustration, our discussion is in terms of standardized predictors satisfying $E(\mathbf{X}) = 0$ and $\text{Cov}(\mathbf{X}) = \mathbf{I}_p$. Here we present the rationale of the UIF method that is in spirit a unification of forward regression, from a viewpoint of Stein's 1981 lemma, and inverse regression, from a viewpoint of the Fourier method in Zhu and Zeng (2006).

Note that the conditional independence model (1.1) is equivalent to, for all $\mathbf{t} \in \mathcal{R}^q$,

$$E[e^{i\mathbf{t}^\tau\mathbf{Y}}|\mathbf{X}] = \psi_{\mathbf{Y}|\mathbf{X}}(\mathbf{t}) = \psi_{\mathbf{Y}|\mathbf{B}^\tau\mathbf{X}}(\mathbf{t}) = E[e^{i\mathbf{t}^\tau\mathbf{Y}}|\mathbf{B}^\tau\mathbf{X}]. \tag{2.1}$$

This, taking the form of the conditional mean, relates to the CS rather than the central mean subspace (CMS, Cook and Li (2002)) only. Therefore, it suffices to estimate the first derivative of the conditional characteristic function to recover the CS in terms of

$$\frac{\partial[\psi_{\mathbf{Y}|\mathbf{X}}(\mathbf{t})]}{\partial\mathbf{X}} = \frac{\mathbf{B}\partial[\psi_{\mathbf{Y}|\mathbf{B}^\tau\mathbf{X}}(\mathbf{t})]}{\partial(\mathbf{B}^\tau\mathbf{X})}. \tag{2.2}$$

To estimate the column space spanned by $\partial[\psi_{\mathbf{Y}|\mathbf{X}}(\mathbf{t})]/\partial\mathbf{X}$, Zhu and Zeng (2006) proposed a Fourier method through the kernel matrix

$$\mathbf{M} = E[\mathbf{S}(\boldsymbol{\Omega},\mathbf{T})\overline{\mathbf{S}}(\boldsymbol{\Omega},\mathbf{T})], \tag{2.3}$$

where the expectation is taken with respect to the random variables $\boldsymbol{\Omega}$ and $\mathbf{T}$, and the seed vector

$$\mathbf{S}(\boldsymbol{\omega},\mathbf{t}) = E\Big[e^{i\boldsymbol{\omega}^\tau\mathbf{X}} \times \frac{\partial(\psi_{\mathbf{Y}|\mathbf{X}}(\mathbf{t}))}{\partial\mathbf{X}}\Big] \tag{2.4}$$

is, in spirit, the Fourier expansion of $\partial(\psi_{\mathbf{Y}|\mathbf{X}}(\mathbf{t}))/\partial\mathbf{X} \times f(\mathbf{X})$ in which $f(\cdot)$ is the density function of $\mathbf{X}$. To estimate $\mathbf{M}$, Zhu and Zeng (2006) assumed $f(\mathbf{X})$ to be given; this is restrictive in the high dimensional regression context. When the distribution of the predictors is unknown, Zhu and Zeng's (2006) Fourier method deserves further exploration.

Recall that Lemma 4 in Stein (1981) is applicable to estimating the average of $\partial(\psi_{\mathbf{Y}|\mathbf{X}}(\mathbf{t}))/\partial\mathbf{X}$ when the standardized predictor vector $\mathbf{X}$ is assumed to be normally distributed. From a forward regression viewpoint, a direct application of the Lemma has

$$E\Big[\frac{\partial[\psi_{\mathbf{Y}|\mathbf{X}}(\mathbf{t})]}{\partial\mathbf{X}}\Big] = E[\mathbf{X}E(e^{i\mathbf{t}^\tau\mathbf{Y}}|\mathbf{X})] = E(\mathbf{X}e^{i\mathbf{t}^\tau\mathbf{Y}}) =: \boldsymbol{\phi}(\mathbf{t}). \tag{2.5}$$

In other words, Stein's Lemma, together with (2.2), shows that the real and imaginary part of $\boldsymbol{\phi}(\mathbf{t})$, denoted by $\boldsymbol{\alpha}(\mathbf{t})$ and $\boldsymbol{\beta}(\mathbf{t})$ respectively, satisfy span$\{\boldsymbol{\alpha}(\mathbf{t}),\boldsymbol{\beta}(\mathbf{t})\}$ $\subseteq$ span$(\mathbf{B}) = \mathcal{S}_{\mathbf{Y}|\mathbf{X}}$, for any given $\mathbf{t}\in\mathcal{R}^q$.

The seed vector $\boldsymbol{\phi}$ can also be defined from an inverse regression perspective. Similar to Zhu and Zeng's (2006) method, $\boldsymbol{\phi}(\mathbf{t}) = E[e^{i\mathbf{t}^\tau\mathbf{Y}}E(\mathbf{X}|\mathbf{Y})]$ can be regarded as the Fourier expansion for the inverse regression $E(\mathbf{X}|\mathbf{Y})\times g(\mathbf{Y})$ where $g(\mathbf{Y})$ is the density function of $\mathbf{Y}$. However, the density function of $\mathbf{Y}$ does not appear in $\boldsymbol{\phi}(\mathbf{t})$ by the law of iterated expectations, and thus we avoid assuming a parametric form for $f(\mathbf{X})$ or $g(\mathbf{Y})$; this differs from Zhu and Zeng's (2006) Fourier method. Here we assume the widely used *linearity condition.* The result is stated as follows.

Proposition 1. *Assume that* $E(\mathbf{X}|\mathbf{B}^\tau\mathbf{X}) = P_{\mathbf{B}}\mathbf{X}$. *Then span*$\{\boldsymbol{\alpha}(\mathbf{t}),\boldsymbol{\beta}(\mathbf{t})\}\subseteq \mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ *for any fixed* $\mathbf{t}\in\mathcal{R}^q$.

The *linearity condition* is typically regarded as mild. Hall and Li (1993) showed that, if the structural dimension K of CS remains fixed as the dimension of the predictors p increases, this condition holds to a good approximation in many problems. See also Diaconis and Freedman (1984) and Cook and Ni (2005).

This proposition implies that $\boldsymbol{\phi}(\mathbf{t})$ for any given $\mathbf{t}$ helps to infer CS. Then we define $\mathbf{M}$ as in (1.3). Clearly, the fact $\{\mathbf{v}\perp\boldsymbol{\phi}(\mathbf{t}),$ for all $\mathbf{t}\in\mathcal{R}^q\}$, which means that $\{\mathbf{v}^\tau\boldsymbol{\alpha}(\mathbf{t})=0,$ and $\mathbf{v}^\tau\boldsymbol{\beta}(\mathbf{t})=0,$ for all $\mathbf{t}\in\mathcal{R}^q\}$, is equivalent to the fact $\{\mathbf{v}\perp\mathbf{M}\}$ if $\mathbf{T}$ be a q-dimensional random vector whose support is $\mathcal{R}^q$. This motivates the following proposition.

Proposition 2. *In addition to the linearity condition, assume that* $\mathbf{T}$ *is supported on* $\mathcal{R}^q$. *Then* $\mathbf{M}\subseteq\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$.

It is worth pointing out that $\mathbf{M}$ still works even when the response is discrete or categorial, as is illustrated by an example once used in Li, Zha, and Chiaromonte (2005).

Example 1. Let $\mathbf{Y}\sim$ Bernoulli(1/2). Let $\mathbf{X}=(X_1,X_2)^\tau$ satisfy $\mathbf{X}|(\mathbf{Y}=i)\sim N(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$, with $\boldsymbol{\mu}_0=(0,-1/\sqrt{2})^\tau$ and $\boldsymbol{\mu}_1=(0,1/\sqrt{2})^\tau$, and $\boldsymbol{\Sigma}_0=\boldsymbol{\Sigma}_1=$

diag$\{1, 1/2\}$, where diag$\{a_1, \ldots, a_p\}$ denotes a diagonal matrix with a_i as the i-th entry. Here $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ is one-dimensional span$\{(0, 1)^\tau\}$. To verify the efficacy of the UIF$_1$ method, we calculate $\mathbf{M}$ as follows. Note that $\boldsymbol{\phi}(\mathbf{t}) = [\boldsymbol{\mu}_0 + e^{i\mathbf{t}}\boldsymbol{\mu}_1]/2$, and $\boldsymbol{\mu}_0\boldsymbol{\mu}_0^\tau = -\boldsymbol{\mu}_0\boldsymbol{\mu}_1^\tau = -\boldsymbol{\mu}_1\boldsymbol{\mu}_0^\tau = \boldsymbol{\mu}_1\boldsymbol{\mu}_1^\tau = \text{diag}\{0, 1/2\}$. Thus, $\mathbf{M} = \boldsymbol{\mu}_0\boldsymbol{\mu}_0^\tau E[1 - \cos(\mathbf{T})]/2$, which implies span$\{\mathbf{M}\} \subseteq \mathcal{S}_{\mathbf{Y}|\mathbf{X}}$. This example shows that $\mathbf{M}$ can apply to regressions with discrete numerical response.

While Zhu and Zeng's (2006) Fourier method is designed for regressions with univariate response, we notice that it can be readily extended to handle the multivariate response case. Recall the definition of the seed matrix $\mathbf{S}(\boldsymbol{\omega}, \mathbf{t})$ in (2.4). Following similar arguments as in Zhu and Zeng (2006), we can show that

$$\mathbf{S}(\boldsymbol{\omega}, \mathbf{t}) = -E[(i\boldsymbol{\omega} + \mathbf{G}(\mathbf{X}))e^{i\boldsymbol{\omega}^\tau\mathbf{X} + i\mathbf{t}^\tau\mathbf{Y}}], \tag{2.6}$$

where $\mathbf{G}(\mathbf{X}) = \partial \log f(\mathbf{X})/\partial\mathbf{X}$. If $\mathbf{X}$ is standard normal, we have

$$\mathbf{S}(0, \mathbf{t}) = \boldsymbol{\phi}(\mathbf{t}). \tag{2.7}$$

The connection (2.7) leads to the following result.

Proposition 3. *Under the conditions in Proposition* 2, span$\{\mathbf{M}\}$ = span$\{$Cov $[E(\mathbf{X}|\mathbf{Y})]\}$.

We remark here that Proposition 8 of Zhu and Zeng (2006) achieves the same result, but they assumed the standard normal distribution, while the UIF$_1$ method requires only the linearity condition. Moreover, the UIF$_1$ method avoids estimating $E(\mathbf{X}|\mathbf{Y})$, which is often challenging when the dimension of $\mathbf{Y}$ is large, though it spans the identical space as does SIR.

3. Implementation

3.1. Estimation

Let $\{(\mathbf{x}_i^\tau, \mathbf{y}_i^\tau)^\tau, i = 1, \ldots, n\}$ be a random sample of $(\mathbf{X}^\tau, \mathbf{Y}^\tau)^\tau$, and $\{\mathbf{t}_i, i = 1, \ldots, n\}$ be an independent random sample of $\mathbf{T}$. We illustrate how to choose $\mathbf{T}$ below.

For any given $\mathbf{t}$ the estimate of $\boldsymbol{\psi}(\mathbf{t})$, denoted by $\boldsymbol{\psi}_n(\mathbf{t})$, is a classical moment estimator. Specifically,

$$\boldsymbol{\psi}_n(\mathbf{t}) = \frac{1}{n}\sum_{j=1}^{n} e^{i\mathbf{t}^\tau\mathbf{y}_j}\mathbf{x}_j = \frac{1}{n}\sum_{j=1}^{n}[\cos(\mathbf{t}^\tau\mathbf{y}_j) + i\sin(\mathbf{t}^\tau\mathbf{y}_j)]\mathbf{x}_j = \boldsymbol{\alpha}_n(\mathbf{t}) + i\boldsymbol{\beta}_n(\mathbf{t}),$$

and thus the estimate of $\mathbf{M}$, written as $\mathbf{M}_n$, takes the form

$$\mathbf{M}_n = \frac{1}{n}\sum_{k=1}^{n}\text{real}\{\boldsymbol{\psi}_n(\mathbf{t}_k)\bar{\boldsymbol{\psi}}_n^\tau(\mathbf{t}_k)\} = \frac{1}{n}\sum_{k=1}^{n}[\boldsymbol{\alpha}_n(\mathbf{t}_k)\boldsymbol{\alpha}_n^\tau(\mathbf{t}_k) + \boldsymbol{\beta}_n(\mathbf{t}_k)\boldsymbol{\beta}_n^\tau(\mathbf{t}_k)]. \tag{3.1}$$

3.2. Choice of T

In the population version, we prefer a $\mathbf{T}$ whose support is $\mathcal{R}^q$ to pool all seed vectors $\boldsymbol{\phi}(\mathbf{t})$ together to recover CS. However, when $\|\mathbf{T}\|$ is too large, a relatively large amount of weight gets assigned to patterns with high frequencies which being sensitive to noise can make our method unstable (Zhu and Zeng (2006)). Our main idea for selecting $\mathbf{T} = (\mathbf{t}_1, \ldots, \mathbf{t}_q)^\tau$ is illustrated as follows. To ensure that the support of $\mathbf{T}$ is $\mathcal{R}^q$, we choose $\mathbf{t}_i$'s, *i.i.d.* $N(0, \sigma^2)$ and, it remains to specify the parameter σ^2. We remark here that the normality is not essential, as we shall see later.

Recall the definitions of $\boldsymbol{\phi}$ in (1.2) and $\mathbf{M}$ in (1.3). Notice that $e^{i\cdot\mathbf{T}^\tau\mathbf{Y}} = e^{i\cdot(\mathbf{T}^\tau\mathbf{Y}+2\pi)}$, which implies that when $|\mathbf{T}^\tau\mathbf{Y}| > \pi$, $\mathbf{T}^\tau\mathbf{Y}$ does not provide additional information in recovering CS. To ensure estimation efficacy, we can simply choose $\mathbf{T}$ satisfying $|\mathbf{T}^\tau\mathbf{Y}| \leq \pi$ with large probability $(1-s)$, namely,

$$P\{|\mathbf{T}^\tau\mathbf{Y}| > \pi\} \leq s. \tag{3.2}$$

By the Chebyshev inequality, the LHS of (3.2) is less than or equal to $\mathrm{Var}(\mathbf{T}^\tau\mathbf{Y})/\pi^2$. Accordingly, (3.2) is reduced to finding σ^2 to satisfy

$$\mathrm{Var}(\mathbf{T}^\tau\mathbf{Y}) = E(\mathbf{Y}^\tau\mathbf{Y})\sigma^2 \leq s\pi^2. \tag{3.3}$$

The equality follows from the fact that $\mathbf{T}\perp\!\!\!\perp\mathbf{Y}$. For any given s, we take

$$\sigma^2 \leq \frac{s\pi^2}{E(\mathbf{Y}^\tau\mathbf{Y})}. \tag{3.4}$$

Still, if σ^2 is too small (e.g. $\|\mathbf{T}\|$ is close to 0), then $\boldsymbol{\phi}(\mathbf{t})$ is close to zero. Consequently, the column space of $\mathbf{M}$ is close to the null space, and will miss some interesting directions in CS. We prefer the largest σ^2 (3.4) and so choose $\sigma^2 = s\pi^2/E(\mathbf{Y}^\tau\mathbf{Y})$.

3.3. UIF$_1$ algorithm

We summarize the algorithm of the UIF$_1$ method below.

Step 1 Standardize the predictors to have zero mean and identity covariance matrix.

Step 2 Specify the probability s, and estimate $E(\mathbf{Y}^\tau\mathbf{Y})$ by $\sum_{i=1}^n \mathbf{y}_i^\tau\mathbf{y}_i/n$. Randomly sample $t_{k,j}$'s, for $k = 1, \ldots, n$, and $j = 1, \ldots, q$, from $N(0, \widehat{\sigma}^2)$ with $\widehat{\sigma}^2 = n \cdot s \cdot \pi^2 / \sum_{i=1}^n \mathbf{y}_i^\tau\mathbf{y}_i$. (We can usually choose a small value from $0.02 < s < 0.30$. Our limited experience shows that the performance of UIF$_1$ is insensitive to the choice s. Throughout our empirical studies, we set $s = 0.10$.)

Step 3 Calculate $\mathbf{M}_n$ from (3.1), where $\mathbf{t}_k = (t_{k,1}, \ldots, t_{k,q})^\tau$.

Step 4 Find the spectral decomposition of $\mathbf{M}_n$. The eigenvectors associated with the largest K eigenvalues of $\mathbf{M}$ are used to estimate the CS.

4. Simulation

The performance of our method is evidenced here through synthetic examples. We also develop the method UIF$_2$ as a subsequent development.

The seven competitors that are designed to recover CS for multivariate response data are used to compare with our methods.

1: The multi-dimensional slicing method (MS), which is the direct extension of the methodology designed for scalar Y. That is , slicing the multi-dimensional $\mathbf{Y}$ into hypercubes, just as one slices the scalar Y into intervals in the original form of SIR or SAVE. In our simulation, we followed the standard slicing scheme: the first slice was made on the first response, yielding h slices of equal length, each of which was then further sliced into h sub-slices according to the second variable, and so on. We took $h = 3, 4, 5, 6$, and 7, corresponding to sample size $n = 100, 200, 400, 800$ and $1,600$ respectively.

2: Alternating SIR (aSIR, Li, Zha, and Chiaromonte (2005)), in which case we slice the M.P. variates $\mathbf{M}_1^\tau \mathbf{Y}$ instead of slicing the entire $\mathbf{Y}$. As the dimension of $\mathbf{M}_1^\tau \mathbf{Y}$ is lower than that of $\mathbf{Y}$, aSIR can improve the accuracy of MS. The initial values of $\mathbf{M}_1$ are the canonical directions identified by $\max_{\boldsymbol{\theta},\boldsymbol{\beta}} \rho(\boldsymbol{\theta}^\tau \mathbf{Y}, \boldsymbol{\beta}^\tau \mathbf{X})$.

3: Nearest neighbor inverse regression (nnIR, Hsing (1999)), in which the slices are determined by nearest neighbors.

4: The K-means inverse regressions (Setodji and Cook (2004)), including KSIR and KSAVE algorithms, which are the same as SIR and SAVE except that the slices are replaced by K-means clusters. In our simulations, we dropped those clusters which contained just one point because KSIR and KSAVE are sensitive to outliers.

5: The methods based on the estimation of central moment spaces that were proposed by Yin and Bura (2006). We only consider two of them, based on the matrices $\mathbf{K}_{21c}$ and $\mathbf{K}_{22c}$ in their paper. The first of these targets the space spanned by the first and the second central moment subspace

$$\operatorname{span}\left(\mathcal{S}_{E(\mathbf{Y}|\mathbf{X})}, \mathcal{S}_{E(\mathbf{Y}^{\otimes 2}|\mathbf{X})}\right),$$

and the second targets the central mean space $\mathcal{S}_{E(\mathbf{Y}|\mathbf{X})}$. We refer to the methods as YB21 and YB22, respectively.

6: The marginal combination method (MC), which recovers the joint CS $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ by pooling together the marginal dimensional reduction subspaces $\mathcal{S}_{Y_i|\mathbf{X}}, i =$

$1, \cdots, q$. We remark here that MC is a special case of the following projective resampling method.

7: The projective resampling method (Li, Wen, and Zhu (2008) is based on the following simple fact: suppose, for any $\mathbf{t} \in \mathcal{R}^q$, $\mathbf{M}(\mathbf{t})$ is a $p \times p$ matrix whose column space spans $\mathcal{S}_{\mathbf{t}^\tau \mathbf{Y}|\mathbf{X}}$. Note that $\mathbf{t}^\tau \mathbf{Y}$ is scalar, thus $\mathcal{S}_{\mathbf{t}^\tau \mathbf{Y}|\mathbf{X}}$ can be easily recovered by many existing methods such as SIR, SAVE, and Directional Regression. In our simulation, the projected CS $\mathcal{S}_{\mathbf{t}^\tau \mathbf{Y}|\mathbf{X}}$ was estimated with SIR, SAVE, or the Directional Regression algorithm, and the corresponding methods are referred to as PSIR, PSAVE and PDR. Let $\mathbf{T}$ be random vector defined on the unit sphere in $\mathcal{R}^q$. Then, $\text{span}\{E[\mathbf{M}(\mathbf{T})]\} = \mathcal{S}_{\mathbf{Y}|\mathbf{X}}$. The LHS can be estimated through Monte Carlo simulation.

We use the trace correlation coefficient $r(K) = \text{trace}(P_{\mathbf{B}} P_{\widehat{\mathbf{B}}})/K$ proposed in Ferré (1998) to measure the closeness of CS $\text{span}\{\mathbf{B}\}$ and its estimate $\text{span}\{\widehat{\mathbf{B}}\}$, where K is the structural dimensionality. It can be verified that $0 \leq r(K) \leq 1$, and larger $r(K)$ values indicate better performance. In particular, $r(K) = 1$ if $\mathcal{S}(\widehat{\mathbf{B}})$ is identical to $\mathcal{S}(\mathbf{B})$ and 0 if $\mathcal{S}(\widehat{\mathbf{B}})$ is perpendicular to $\mathcal{S}(\mathbf{B})$.

Throughout we take $\mathbf{X}_{20\times 1}$ to be normally distributed with mean zero and identity covariance matrix. The dimension K of CS is assumed to be known; the mean and standard deviation of $r(K)$ reported in the following tables were computed from 1,000 repetitions. Unless stated otherwise, in PSIR and KSIR, the slice numbers or cluster numbers were 10, 10, 20, 40, and 40 corresponding to the sample size $n = 100, 200, 400, 800$, and $1,600$. Li (1991) and Setodji and Cook (2004) showed that the results of the SIR-based methods were not very sensitive to the number of slices or clusters. We chose $s = 0.10$ throughout our empirical studies.

Example 2. Consider the simple model

$$\mathbf{Y}|\mathbf{X} \sim N_2(\mathbf{0},\ \mathbf{\Sigma}(\boldsymbol{\beta}^\tau \mathbf{X})),$$

$$\mathbf{\Sigma}(\boldsymbol{\beta}^\tau \mathbf{X}) = \begin{pmatrix} 1 & \sin(\boldsymbol{\beta}^\tau \mathbf{X}) \\ \sin(\boldsymbol{\beta}^\tau \mathbf{X}) & 1 \end{pmatrix},$$

and $\boldsymbol{\beta} = (0.8, 0.6, 0, 0, \cdots, 0, 0)^\tau$. Clearly, $\mathcal{S}_{\mathbf{Y}|\mathbf{X}} = \text{span}\{\boldsymbol{\beta}\}$. However, as $\mathcal{S}_{Y_i|\mathbf{X}} = \{\mathbf{0}\}$ for $i = 1, 2$, we do not expect MC to work well. As $\text{Cov}(\mathbf{X}, \mathbf{Y}) = E(\mathbf{X}\mathbf{Y}) = \mathbf{0}$, we cannot get the desired initial value of $\mathbf{M}$ efficiently in aSIR and the method does not work. The comparison of the performance of the other estimators is presented in Table 4.1. We can see that UIF_1 performed better than all other methods when the sample size was comparatively small. When n was large, the results were very close except for the MC and aSIR methods. In all UIF_1 was effective in detecting the true direction.

Table 4.1. Comparison based on Example 2.

	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 1,600$
MC	0.0540(0.0712)	0.0479(0.0594)	0.0538(0.0708)	0.0488(0.0648)	0.0533(0.0690)
MS	0.2674(0.1831)	0.6027(0.1686)	0.8096(0.1672)	0.9203(0.0274)	0.9636(0.0125)
aSIR	0.0540(0.0820)	0.0545(0.0796)	0.0847(0.1495)	0.0848(0.1602)	0.0878(0.1793)
nnSIR	0.2335(0.1899)	0.4134(0.2266)	0.6593(0.0484)	0.8387(0.0670)	0.9194(0.0309)
KSIR	0.3816(0.2016)	0.6787(0.1441)	0.8622(0.0568)	0.9382(0.0204)	0.9697(0.0104)
YB21	0.4566(0.1514)	0.6801(0.1032)	0.8257(0.0732)	0.9095(0.0294)	0.9524(0.0151)
PSIR	0.2653(0.2050)	0.6343(0.1867)	0.8678(0.0499)	0.9390(0.0212)	0.9716(0.0098)
PDR	0.1230(0.1294)	0.1463(0.1536)	0.2949(0.2111)	0.5901(0.2005)	0.8463(0.0660)
UIF$_1$	0.3796(0.2179)	0.7114(0.1355)	0.8733(0.0457)	0.9379(0.0210)	0.9682(0.0110)

Table 4.2. Comparison based on Example 3.

	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 1,600$
MC	0.5189(0.1117)	0.7899(0.0762)	0.8984(0.0321)	0.9627(0.0106)	0.9820(0.0048)
MS	0.3935(0.1270)	0.5907(0.1243)	0.7904(0.1009)	0.9246(0.0291)	0.9728(0.0090)
aSIR	0.6125(0.1125)	0.8296(0.0663)	0.9289(0.0214)	0.9652(0.0108)	0.9833(0.0049)
nnSIR	0.5126(0.1163)	0.7086(0.1159)	0.8713(0.0487)	0.9428(0.0180)	0.9737(0.0080)
KSIR	0.7126(0.0854)	0.8663(0.0416)	0.9415(0.0168)	0.9722(0.0078)	0.9879(0.0035)
YB21	0.4252(0.1257)	0.6241(0.1322)	0.8077(0.1043)	0.9278(0.0302)	0.9673(0.0105)
PSIR	0.7984(0.0605)	0.8970(0.0313)	0.9500(0.0145)	0.9752(0.0070)	0.9880(0.0034)
PDR	0.6018(0.1180)	0.8003(0.0725)	0.9253(0.0252)	0.9654(0.0102)	0.9830(0.0046)
UIF$_1$	0.7751(0.0627)	0.8892(0.0356)	0.9480(0.0175)	0.9707(0.0081)	0.9844(0.0043)

Example 3. We consider a slightly more complicated model in which $Y_1 = 1 + X_1 + \sin(X_2 + X_3) + \varepsilon_1$, $Y_2 = (X_2 + X_3)/(0.5 + (X_1 + 1)^2) + \varepsilon_2$, $Y_3 = |X_1|\varepsilon_3$, $Y_4 = \varepsilon_4$, $Y_5 = \varepsilon_5$, where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_5)^\tau \sim N_5(\mathbf{0}, \mathbf{\Sigma})$ with

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & -\frac{1}{2} & 0 & 0 & 0 \\ -\frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{4} \end{pmatrix}.$$

In this model, CS is spanned by directions $\boldsymbol{\beta}_1 = (1, 0, 0, \cdots, 0, 0)^\tau$ and $\boldsymbol{\beta}_2 = (0, 1, 1, 0, 0, \cdots, 0, 0)^\tau$. We can see from Table 4.2 that all methods converged for large sample size and recovered CS. The performance of UIF$_1$ was comparable to PSIR, though it outperformed other methods significantly when the sample size was small. The aSIR method performed better than MS in this example, this

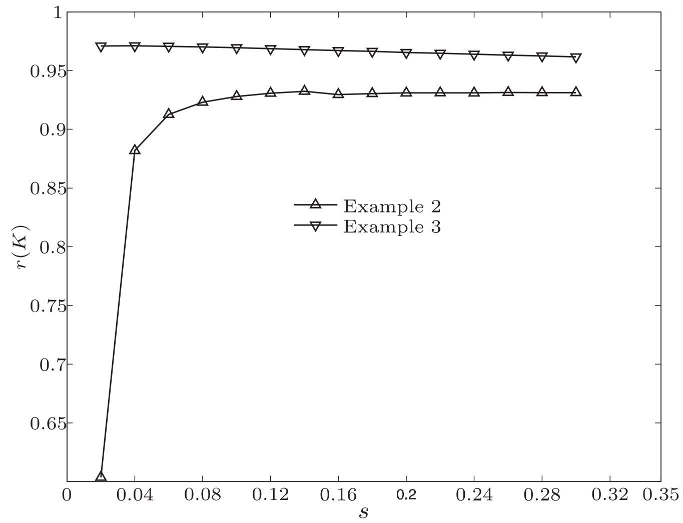

Figure 4.1. The average of the empirical $r(K)$ values versus different s values. We fixed the sample size n=800 in Examples **2** and **3**.

is because only Y_1, Y_2 and Y_3 depend on $\mathbf{X}$, so slicing on Y_4 and Y_5 is useless, the slicing on the reduced M.P. variates $\mathbf{M}_1^\tau \mathbf{Y}$ can recover as many directions as slicing the entire $\mathbf{Y}$.

Remark. We also made simulations in Examples **2** and **3** to show our UIF$_1$ method insensitive to the choice of s by applying UIF$_1$ with different s values to recover $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$. The average of the empirical $r(K)$ values over 1,000 repetitions versus different s values with a fixed sample size $n = 800$ are shown in Figure 4.1; clearly, for $0.02 < s < 0.30$, our UIF$_1$ method had a stable and satisfactory performance, which suggests that UIF$_1$ is robust to the choice of s values.

Example 4. The SIR-based algorithms such as KSIR and PSIR fail to detect directions when the link function is even and symmetric about the functional components, as in the following model. Meanwhile, Yin and Bura (2006) argued that YB21 is analogous to SIR methodology and also fails to detect such structure. To counter this, they proposed the YB22 method. Our method UIF$_1$ suffers from the same problem as well, as can be easily seen from an inverse regression perspective. Thus we introduce a second moment based method, called UIF$_2$, that replaces $\boldsymbol{\phi}(\mathbf{t}) = E(e^{i\mathbf{t}^\tau \mathbf{Y}}\mathbf{X})$ at (1.2) by $E[e^{i\mathbf{t}^\tau \mathbf{Y}}(\mathbf{X}\mathbf{X}^\tau - \mathbf{I}_p)]$, and we construct $\mathbf{M}$ in the same way as in (1.3). The UIF$_2$ method is essentially similar to the UIF$_1$ method that targets the CS under *the linearity condition* and *the constant variance condition* (Cook and Weisberg (1991)), and hence the theoretical investigations are omitted in the present context.

Table 4.3. Comparison based on Example 4.

$n = 200$	YB21	0.2689 (0.1240)	UIF$_1$	0.1499 (0.0855)	
	YB22	0.5369 (0.0984)	UIF$_2$	0.7332 (0.0801)	
	$H = 2$	$H = 5$	$H = 10$	$H = 20$	$H = 40$
KSAVE	0.3287(0.0803)	0.3170 (0.1040)	0.3220(0.1209)	0.3110(0.1307)	0.2095(0.1182)
PSAVE	0.4002(0.1169)	0.6472 (0.1216)	0.5629(0.1235)	0.4319(0.1177)	0.3025(0.1128)
$n = 400$	YB21	0.2788(0.1261)	UIF$_1$	0.1482(0.0855)	
	YB22	0.6452(0.0960)	UIF$_2$	0.8988(0.0288)	
	$H = 2$	$H = 5$	$H = 10$	$H = 20$	$H = 40$
KSAVE	0.4025(0.0568)	0.3861(0.0992)	0.3905(0.1246)	0.4434(0.1574)	0.3822(0.1506)
PSAVE	0.5841(0.1077)	0.8738(0.0431)	0.8332(0.0618)	0.7471(0.0997)	0.5878(0.1244)
$n = 800$	YB21	0.2996(0.1358)	UIF$_1$	0.1532(0.0862)	
	YB22	0.7501(0.0984)	UIF$_2$	0.9529(0.0116)	
	$H = 2$	$H = 5$	$H = 10$	$H = 20$	$H = 40$
KSAVE	0.4571(0.0399)	0.4592(0.0983)	0.4412(0.1115)	0.5575(0.1811)	0.5763(0.1770)
PSAVE	0.7599(0.0974)	0.9490(0.0151)	0.9368(0.0196)	0.9177(0.0272)	0.8775(0.0480)

In the example, we compare UIF$_2$ with YB22 and the SAVE-based methods KSAVE and PSAVE. The model takes $Y_1 = 1 + X_1^2 + \sin(X_2 + 3X_3) + \varepsilon_1,\ Y_2 = X_1(X_2 + 3X_3) + (1 + (X_2 + 3X_3)^2)\varepsilon_2,\ Y_3 = \varepsilon_3,\ Y_4 = \varepsilon_4,\ Y_5 = \varepsilon_5$, where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_5)^{\tau} \sim N_5(\mathbf{0}, \boldsymbol{\Sigma})$ with the $\boldsymbol{\Sigma}$ given in the previous example. The CS is spanned by directions $\boldsymbol{\beta}_1 = (1, 0, 0, \cdots, 0, 0)^{\tau}$ and $\boldsymbol{\beta}_2 = (0, 1, 3, 0, 0, \cdots, 0, 0)^{\tau}$. Li and Zhu (2007) showed that the slice number is crucial to SAVE, we report the results of KSAVE and PSAVE with different slice/cluster numbers here. We can see that the results of KSAVE were not satisfactory if the slice number was not chosen properly, even when sample size was large. YB22 need not choose the slice number, but it converged very slowly. To be precise, the result for the YB22 method was 0.8596(0.0564) when $n = 1,600$, and 0.9230(0.0334) when $n = 3,200$. Let H denote the number of slices in PSAVE or the number the clusters in KSAVE. As can be seen from Table 4.3, the PSAVE method was efficient if the slice number was properly selected. Comparatively, UIF$_2$ performed much better, and moreover, it avoided selection of the slice number.

Example 5. Now we consider a model including both linear structure and a symmetric component. Suppose $Y_1 = 1 + X_1^2 + e^{X_2} \cdot \varepsilon_1,\ Y_2 = X_1(X_1 + X_2) + \varepsilon_2,\ Y_3 = X_1^2 + 0.5 \cdot \varepsilon_3,\ Y_4 = X_2^2 + 0.5 \cdot \varepsilon_4,\ Y_5 = \varepsilon_5$, where ε_i's are independent standard normal variables. The central space is spanned by $\boldsymbol{\beta}_1 = (1, 0, 0, \cdots, 0, 0)^{\tau}$ and

Table 4.4. Comparison based on Example 5.

	$n = 200$	$n = 400$	$n = 800$	$n = 1,600$
MC	0.3088(0.1009)	0.4198(0.0724)	0.4785 (0.0503)	0.5116(0.0494)
MS	0.5689(0.1381)	0.6946(0.1184)	0.7862 (0.0895)	0.8409(0.0672)
nnSIR	0.3913(0.1107)	0.5079(0.0943)	0.6063 (0.1210)	0.7264(0.1315)
KSIR	0.4567(0.0987)	0.5552(0.1073)	0.6737 (0.1289)	0.7905(0.1122)
KSAVE	0.3614(0.1927)	0.4707(0.2085)	0.5772 (0.2313)	0.7054(0.2392)
YB21	0.4240(0.1151)	0.4935(0.1195)	0.5344 (0.1193)	0.5676(0.1132)
YB22	0.8153(0.0624)	0.9171(0.0302)	0.9623 (0.0145)	0.9825(0.0053)
PSIR	0.3824(0.0950)	0.4873(0.0806)	0.5503 (0.0924)	0.5854(0.1012)
PSAVE	0.8656(0.0447)	0.9509(0.0128)	0.9787 (0.0051)	0.9878(0.0024)
PDR	0.7014(0.1055)	0.9145(0.0240)	0.9514 (0.0131)	0.9805(0.0047)
UIF1	0.3942(0.0967)	0.4788(0.0756)	0.5285 (0.0761)	0.5601(0.0803)
UIF2	0.8468(0.0505)	0.9510(0.0120)	0.9748 (0.0060)	0.9882(0.0029)

$\boldsymbol{\beta}_2 = (0, 1, 0, 0, \cdots, 0, 0)^{\tau}$. Based on our experience in the previous example, the slice number in the PSAVE algorithm was fixed at 5, the slice numbers or cluster numbers in KSIR, KSAVE, PSIR and PDR algorithm were taken to be 10, 20, 40, and 40, corresponding to sample sizes 200, 400, 800, and 1,600. The results reported in Table 4.4 show that PSAVE and UIF_2 outperformed other methods, the SIR-based methods were not effective in detecting such model structure.

5. Determinants of Plasma Retinol and Beta-carotene Levels

Observational studies suggested that low dietary intake or low plasma concentrations of retinol, beta-carotene, or other carotenoids might be associated with increased risk of developing certain types of cancer. However, relatively few studies have investigated the determinants of plasma concentrations of these micronutrients. To address this issue, Nierenberg et al. (1989) designed a cross-sectional study to investigate the associations of retinol and beta-carotene plasma concentrations with 12 personal characteristics and dietary factors in 315 patients with nonmelanoma skin cancer, enrolled at four American study centers in a cancer prevention clinical trial.

The response $\mathbf{Y} = (Y_1, Y_2)$ is two-dimensional, where Y_1 and Y_2 denote, respectively, the plasma beta-carotene (ng/ml) and the plasma retinol (ng/ml). The quantitative predictors $\mathbf{X} = (X_1, \ldots, X_9)^T$ are related to personal characteristics and dietary factors: the age (in years) of each subject X_1, the quetelet variable defined by weight/height^2 X_2 (kg/m^2), the calories consumed per day X_3, fat consumed per day X_4 ($grams$), fiber consumed per day X_5 ($grams$),

Table 5.5. Comparison of different candidate methods.

competitors	MC	MS	PDR	UIF$_1$	PSIR	YB21	YB22
variability	0.2189	0.2391	0.0788	0.0669	0.1036	0.1189	0.0741

number of alcoholic drinks consumed per week X_6, cholesterol consumed (mg) per day X_7, dietary beta-carotene consumed (mcg) per day X_8, and dietary retinol consumed (mcg) per day X_9. Categorical predictors such as gender, vitamin use, and smoking status may be relevant, but are excluded from our analysis to ensure the linearity condition. There was one extremely high leverage point (NO 62) in alcohol consumption (X_6) that was deleted prior to the analysis to avoid possible modeling bias. We further standardized the predictors to have zero mean and identity covariance matrix before our analysis.

The multiple correlation coefficient between plasma beta-carotene levels and the full linear model was 0.15, and the multiple correlation coefficient between the plasma retinol and the full linear model was only 0.11. The linear model may be insufficient and model-free dimension reduction method is sought. We apply the UIF$_1$ method. The largest two eigenvalues account for over 99% variability, which complies with the BIC type criterion with $\ln(n)$ as the penalty (Zhu and Zhu (2007)) that the joint CS is two-dimensional. In the sequel we assume the structural dimension of the joint CS is known.

Now we compare the performance of UIF$_1$ with some competitors designed for multivariate response data, including the marginal combination method (MC), the multi-dimensional slicing method (MS), the projective resampling method combined with SIR (PSIR), and directional regression (PDR), the methods proposed by Yin and Bura (2006, YB21, YB22). One can refer to Section 4 for detailed descriptions of these methods. To select among dimension reduction estimators $\widehat{\mathcal{S}}_{\mathbf{Y}|\mathbf{X}}$, we follow the idea of Ye and Weiss (2003, p. 972) and prefer the one with smallest variability, assuming no or minimal bias. To put this into practice, we used the bootstrap method to generate a set of $\widehat{\mathcal{S}}^b_{\mathbf{Y}|X}$, and calculated the trace correlation coefficient $r^b(K)$ (Ferré (1998)) between $\widehat{\mathcal{S}}_{\mathbf{Y}|X}$ and $\widehat{\mathcal{S}}^b_{\mathbf{Y}|X}, b = 1, \ldots, B$. We adopted the median of $(1 - r^b(K)), b = 1, \ldots, B$, as a measure of variability of the estimator $\widehat{\mathcal{S}}_{\mathbf{Y}|X}$. The medians for different dimension reduction estimators from 2,000 repetitions between $\widehat{\mathcal{S}}_{\mathbf{Y}|\mathbf{X}}$ and the bootstrap estimator $\widehat{\mathcal{S}}^b_{\mathbf{Y}|\mathbf{X}}$ are summarized in Table 5.5.

In terms of variability, UIF$_1$ performed the best, followed by YB22 and PDR. It is not surprising to find that these methods bested MC and MS.

The coefficients of the first two UIF$_1$ directions are presented in Table 5.6. Recall that the predictors are standardized. If these predictors are independent,

Table 5.6. The coefficients of the first two UIF$_1$ directions.

coefficients	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
1st direction	0.6054	-0.2700	-0.0982	-0.3354	0.0655	0.4344	-0.2940	0.3623	-0.1667
2nd direction	0.3136	0.4860	0.0056	-0.0253	-0.4321	0.4460	0.2193	-0.4792	-0.0370

then we can judge the predictors' contribution by the magnitude of the coefficients. In particular, X_1 may be the most important predictor, while X_3 may be the least important, because it seems irrelevant to $\mathbf{Y}$. We conclude that there is wide variability in plasma concentrations of these micronutrients in humans, and that much of this variability is due to the dietary habits and most personal characteristics, except for the calories consumed each day.

Acknowledgement

We thank two referees, an Associate Editor, and the Editors for their prompt and extremely helpful reviews which, for example, led us to consider a rigorous way to select the random projections. Li-Ping Zhu's work was supported by a NSF grant from National Natural Science Foundation of China (No. 10701035) and the ChenGuang project of Shanghai Education Development Foundation (No. 2007CG33). Li-Xing Zhu's research was partially supported by a grant (HKU 7058/05P) from the Research Grants Council of Hong Kong, and a FRG grant from Hong Kong Baptist University, Hong Kong.

Appendix

All technical derivations are relegated to this Appendix.

Proof of Proposition 1. By the definition of $\boldsymbol{\phi}$, we have that

$$\begin{aligned}\boldsymbol{\phi}(\mathbf{t}) &= E[E(e^{i\mathbf{t}^\tau\mathbf{Y}}\mathbf{X}|\mathbf{B}^\tau\mathbf{X})]\\ &= E[E(e^{i\mathbf{t}^\tau\mathbf{Y}}|\mathbf{B}^\tau\mathbf{X})E(\mathbf{X}|\mathbf{B}^\tau\mathbf{X})]\\ &= \mathbf{B}(\mathbf{B}^\tau\mathbf{B})^{-1}\mathbf{B}^\tau E[E(e^{i\mathbf{t}^\tau\mathbf{Y}}|\mathbf{B}^\tau\mathbf{X})\mathbf{X}] = \mathbf{B}(\mathbf{B}^\tau\mathbf{B})^{-1}\mathbf{B}^\tau\boldsymbol{\phi}(\mathbf{t}).\end{aligned}$$

The second equality holds by the conditional independence (1.1), the third equality follows from the *linearity condition*, and the last equality holds because $E(e^{i\mathbf{t}^\tau\mathbf{Y}}|\mathbf{X}) = E(e^{i\mathbf{t}^\tau\mathbf{Y}}|\mathbf{B}^\tau\mathbf{X})$, which is implied by (1.1). Therefore, the real and imaginary parts of $\boldsymbol{\phi}(\mathbf{t})$ lie in CS. Thus the conclusion follows.

Proof of Proposition 2. Notice that the fact that $\mathbf{v} \perp \boldsymbol{\phi}(\mathbf{t})$, for all $\mathbf{t} \in \mathcal{R}^q$ is equivalent to saying that $\mathbf{v} \perp \mathbf{M}$ if $\mathbf{T}$ is a q-dimensional random vector whose support is $\mathcal{R}^q$. Therefore, the desired conclusion follows directly from Proposition 1 and the linearity condition.

Proof of Proposition 3. We only sketch the outline of this proof because it is parallel to the arguments used in proving Proposition 8 in Zhu and Zeng (2006, p. 1,642). Notice that $\boldsymbol{\phi}$ is in spirit the Fourier transform of $E(\mathbf{X}|\mathbf{Y}=\mathbf{y})f_{\mathbf{Y}}(\mathbf{y})$. Then for any $\mathbf{v} \in \mathcal{R}^p$, $\mathbf{v}^\tau E(\mathbf{X}|\mathbf{Y}=\mathbf{y})f_{\mathbf{Y}}(\mathbf{y}) = 0$ for $\mathbf{y} \in \mathcal{R}^q$ is equivalent to $\mathbf{v}^\tau \boldsymbol{\alpha}(\mathbf{t}) = \mathbf{v}^\tau \boldsymbol{\beta}(\mathbf{t}) = 0$ for all $\mathbf{t} \in \mathcal{R}^q$, where $\boldsymbol{\phi}(\mathbf{t}) = \boldsymbol{\alpha}(\mathbf{t}) + i \cdot \boldsymbol{\beta}(\mathbf{t})$. Hence $\text{span}\{\boldsymbol{\phi}(\mathbf{t}), \mathbf{t} \in \mathcal{R}^p\} = \text{span}\{E(\mathbf{X}|\mathbf{Y}=\mathbf{y}) : \mathbf{y} \in \text{Support}(\mathbf{Y})\}$. The former is exactly the space spanned by $\mathbf{M}$, and the latter is the space aimed at by SIR. Thus the proposition is proved.

References

Aragon, Y. (1997). A Gauss implementation of multivariate sliced inverse regression. *Comput. Statist.* **12**, 355-372.

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

Cook, R. D. and Critchley, F. (2000). Identifying regression outliers and mixtures graphically. *J. Amer. Statist. Assoc.* **95**, 781–794.

Cook, R. D. and Li, B. (2002). Dimension reduction for conditional mean in regression. *Ann. Statist.* **30**, 455-474.

Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *J. Amer. Statist. Assoc.* **100**, 410-428.

Cook, R. D. and Setodji, C. M. (2003). A model-free test for reduced rank in multivariate regression. *J. Amer. Statist. Assoc.* **98**, 340-351.

Cook, R. D. and Weisberg, S. (1991). Discussion of "Sliced inverse regression for dimension reduction". *J. Amer. Statist. Assoc.* **86**, 316-342.

Diaconis, P. and Freedman, D. (1984). Asymptotics of graphical projection pursuit. *Ann. Statist.* **12**, 793-815.

Ferré, L. (1998). Determing the dimension in sliced inverse regression and related methods, *J. Amer. Statist. Assoc.* **93**, 132-140.

Hall, P. and Li, K. C. (1993). On almost linearity of low dimensional projection from high dimensional data. *Ann. Statist.* **21**, 867-889.

Hsing, T. (1999). Nearest neighbor inverse regression. *Ann. Statist.* **27**, 697-731.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.* **86**, 316-342.

Li, K. C., Aragon, Y., Shedden, K., and Agnan, C. T. (2003). Dimension reduction for multivariate response data. *J. Amer. Statist. Assoc.* **98**, 99-109.

Li, B. and Wang, S. L. (2007). On directional regression for dimension reduction. *J. Amer. Statist. Assoc.* **102**, 997-1008.

Li, B., Wen, S. Q., and Zhu L. X. (2008). On a projective resampling method for dimension reduction with multivariate responses. *J. Amer. Statist. Assoc.* **103**, 1177-1186.

Li, B., Zha, H. and Chiaromonte F. (2005). Contour regression: A general approach to dimension reduction. *Ann. Statist.* **33**, 1580-1616.

Li, Y. X. and Zhu, L. X. (2007). Asymptotics for sliced average variance estimation. *Ann. Statist.* **35**, 41-69.

Nierenberg, D. W., Stukel, T. A., Baron, J. A., Dain, B. J. and Greenberg, E. R. (1989). Determinants of plasma levels of beta-carotene and retinol. *Amer. J. Epidemiol.* **130**, 511-521.

Saracco, J. (2005). Asymptotics for pooled marginal slicing estimator based on SIR$_\alpha$ approach. *J. Multi. Anal.* **96**, 117-135.

Setodji, C. M. and Cook, R. D. (2004). K-means inverse regression. *Technometrics* **46**, 421-429.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9**, 1135-1151.

Ye, Z. and Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *J. Amer. Statist. Assoc.* **98**, 968-979.

Yin, X. R. and Bura, E. (2006). Moment based dimension reduction for multivariate response regression. *J. Statist. Plan. Inference* **136**, 3675-3688.

Zhu, L. X. and Fang, K. T. (1996). Asymptotics for the kernel estimates of sliced inverse regression. *Ann. Statist.* **24**, 1053-1067.

Zhu, L. X., Ohtaki, M. and Li, Y. X. (2007). On hybrid methods of inverse refression-based algorithms. *Comput. Statist. Data Anal.* **51**, 2621-2635.

Zhu, L. P. and Yu, Z. (2007). On spline approximation of sliced inverse reression. *Science in China* **50**, 1289-1302.

Zhu, Y. and Zeng, P. (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *J. Amer. Statist. Assoc.* **101**, 1638-1651.

Zhu, L. P. and Zhu, L. X. (2007). On kernel method for sliced average variance estimation. *J. Multi. Anal.* **98**, 970-991.

School of Finance and Statistics, East China Normal University, Shanghai, 200241, China.

E-mail: lpzhu@stat.ecnu.edu.cn

Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Kowloon, Hong Kong.

E-mail: lzhu@math.hkbu.edu.hk

College of Mathematics and Computational Science, Shenzhen University, Shenzhen, Guangdong, 518060, China.

E-mail: sqwen@szu.edu.cn

(Received March 2008; accepted March 2009)