Statistica Sinica **24** (2014), 1195-1214
doi:http://dx.doi.org/10.5705/ss.2012.286

MOST INFORMATIVE COMPONENT ANALYSIS

Yaping Jing, Hong Lei and Yingcun Xia

*University of Electronic Science and Technology of China,
Guizhou University of Finance and Economics and
National University of Singapore*

Abstract: We extend the principal component analysis (PCA) to the investigation of nonlinear dependence among variables, called most informative component analysis (MICA). The most informative components are linear combinations of the variables that capture both linear and nonlinear dependence among the variables. Compared with the existing extensions such as the principal curve and the kernel PCA, MICA is more interpretable and thus more meaningful in statistical analysis. Properties of MICA are investigated, the estimation method is developed, and asymptotics of the estimators are obtained. Data sets are analyzed to illustrate the usefulness of MICA.

Key words and phrases: Dimension reduction, most predictable component, principal component analysis, projection pursuit, unsupervised learning.

1. Introduction

Principal component analysis, developed by Pearson (1901), is a fundamental method in data analysis. It explores the linear dependence in a set of variables $X = (\mathbf{x}_1, \ldots, \mathbf{x}_p)^\top$, and is commonly used to reduce the dimensionality of the dataset by retaining only a few linear combinations of the variables, the principal components (PC), and to extract features from X for better understanding and analysis of the data, such as clustering and pattern recognition. That it is a linear method implies a potential oversimplification of the datasets being analyzed. Some extensions of PCA in a linear framework that focus on different statistical aspects of the data include independent component analysis (Comon (1994)) and the common factor analysis of Pan and Yao (2008).

Extensions of PCA to the investigation of nonlinear dependence amongst the variables have also been considered. Hastie and Stuetzle (1989) proposed the principal curve, and Kramer (1991) proposed an autoassociative neural network (ANN) structure that defines mapping and demapping stages by neural network layers. Schölkopf, Smola, and Müller (1998) proposed a kernel PCA that first maps the original variable set X onto a higher dimensional feature space and then applies PCA to reduce the dimension. By doing so, the nonlinear dependence in

X can be detected. However, the "principal components" in those methods are not easy to interpret because they are neither linear combination nor other simple functions of the original variables. Cook (2007) did a comprehensive review of PCA and proposed an analysis method called principal fitted components (PFC). PFC is calculated "under the supervision" of a response variable Y, and is thus different from PCA as the latter is an unsupervised learning approach. It is known that dimension reduction problems with and without a response variable are quite different.

2. Definition of the Most Informative Component

There are two ways to calculate PCs, one based on the covariance matrix of X, and the other on the correlation coefficient matrix of X. We only consider the latter, equivalent to assuming $\mathrm{Var}(\mathbf{x}_i) = 1$ for $i = 1, \ldots, p$. Let $\Sigma = \mathrm{Var}(X)$ have the eigenvalue-eigenvector decomposition $\Sigma = \Gamma diag(\lambda_1, \ldots, \lambda_p)\Gamma^\top$, where $\Gamma = (\theta_1, \ldots, \theta_p)$ is an orthogonal matrix and $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$. Then $\theta_d^\top X$ is the dth PC, $d = 1, \ldots, p$.

Here is another interpretation of the PCs. For any random vectors $V : p \times 1$ and $U : q \times 1$, define linear conditional expectation as

$$L(V|U) \overset{def}{=} a_0 + b_0^\top U,$$

with

$$(a_0, b_0) = \arg \min_{a:p\times 1, b:q\times p} E\|V - a_0 - b_0^\top U\|^2,$$

where $\|A\| = \{tr(A^\top A)\}^{1/2}$ denotes the Euclidian norm of matrix A. For any fixed $d < p$, the linear combinations $\beta_1^\top X, \ldots, \beta_d^\top X$, or $B^\top X$ where $B = (\beta_1, \ldots, \beta_d)$, that best predict X *linearly* is

$$B_d = \arg\min_B E\|X - L(X|B^\top X)\|^2. \tag{2.1}$$

We call $B_d^\top X$ the first d linearly most informative components (LMIC) of X. The connection between LMICs and the usual principal components is the following.

Proposition 1. *For any $1 \leq d < p$, if $\lambda_d > \lambda_{d+1}$ then the first d principal components and the first d LMICs are in the same space, $\mathcal{S}(B_d) = \mathcal{S}(\theta_1, \ldots, \theta_d)$, where $\mathcal{S}(B)$ denotes the space spanned by the column vectors of B.*

Based on this connection, PCA looks for d $(< p)$ linear combinations of X that are most informative in *linearly* explaining or predicting X. We extend PCA to include nonlinear dependence in X by changing the linear conditional expectation $L(V|U)$ to the general conditional expectation $E(V|U)$. We take the first d most informative components (MIC) of X, denoted by $B^\top X$, such that they minimize

$$E\|X - E(X|B^\top X)\|^2.$$

To simplify this, we investigate an alternative approach by considering one component at a time, similar to the idea of projection pursuit (see for example Huber (1985)). Consider again the traditional PCA. Let $R_0 = X$ and take the first linear most informative component (LMIC), say $\beta_1^\top X$, to minimize

$$E\|R_0 - L(R_0|\beta^\top X)\|^2, \tag{2.2}$$

with respect to β with $\|\beta\| = 1$, and take $R_1 = R_0 - L(R_0|\beta_1^\top X)$. The $(d+1)$th component $\beta_{d+1}^\top X$ is taken to minimize

$$E\|R_d - L(R_d|\beta^\top X)\|^2 \tag{2.3}$$

with respect to β with $\|\beta\| = 1, d \geq 1$.

Proposition 2. *The components defined by* (2.3) *satisfy* $\mathcal{S}(\theta_1, \ldots, \theta_d) = \mathcal{S}(\beta_1, \ldots, \beta_d)$ *for any* $1 \leq d \leq p$ *if* $\lambda_d > \lambda_{d+1}$, $cov(\beta_k^\top X, R_d) = 0$ *for all* $1 \leq k \leq d$, *and* $E\|L(R_k|\beta_{k+1}^\top X)\|^2 = E\|\beta_{k+1}^\top X\|^2$ *for all* $k = 1, \ldots, p-1$.

Motivated by (2.2) and (2.3), we shall make another extension of PCA. The first most informative component (MIC), say $\beta_1^\top X$, is to minimize

$$E\|R_0 - E(R_0|\beta^\top X)\|^2 \tag{2.4}$$

with respect to β with $\|\beta\| = 1, d \geq 1$. The $(d+1)$th MIC, $\beta_{d+1}^\top X$, minimizes

$$E\|R_d - E(R_d|\beta^\top X)\|^2 \tag{2.5}$$

with respect to $\beta : \|\beta\| = 1$. Let $R_{d+1} = R_d - E(R_d|\beta_{d+1}^\top X)$. By repeating this procedure, we can get a sequence of MICs. For convenience, we call $E\|E(R_{d-1}|\beta_d^\top X)\|^2$ the information contained in MIC $\beta_d^\top X$ for $d = 1, \ldots$, which is also the variation or information in R_{d-1} that can be explained by $\beta_d^\top X$.

Proposition 3. *Suppose the first* d *PCs are* $\theta_1^\top X, \ldots, \theta_d^\top X$. *If for any linear combinations* $\ell^\top X$, *there exist vectors* a *and* b *such that* $E(X|\ell^\top X) = a + b\ell^\top X$, *then the first* d *MICs,* $\beta_1^\top X, \ldots, \beta_d^\top X$, *satisfy* $\mathcal{S}(\theta_1, \ldots, \theta_d) = \mathcal{S}(\beta_1, \ldots, \beta_d)$ *for any* $1 \leq d \leq p$ *provided* $\lambda_d > \lambda_{d+1}$. *If the eigenvalues of* $Var(X)$ *distinct,* β_k *and* θ_k *agree up to a sign.*

Some examples are discussed in Section 5. If X is elliptically distributed, then the conditions in Proposition 3 are satisfied; see Cook (2008).

3. Connection with Other Approaches

A PC $\theta_d^\top X$ actually minimizes $E\|R_{d-1} - a_d - c_d\theta_d^\top X\|^2$ with respect to $p \times 1$ vectors a_d, c_d, and θ_d, while a most informative component minimizes $E\|R_d - g_d(\theta_d^\top X)\|^2$, where $g_d(v) = (g_{d1}(v), \dots, g_{dp}(v))^\top$ are unknown link functions. In this sense, MIC is an extension of the principal curve of Hastie and Stuetzle (1989) who considered the case $d = 1$. We proceed in the manner of projection pursuit: if the first approximation is not satisfactory, we consider the second approximation to the remainders of the first approximation by the second component, and continue as needed. The auto-associative model of Girard and Iovleff (2001) has a very similar spirit, but it approximates X by linear combinations of the residuals, and this may not be interpretable statistically.

Wang, Sha, and Jordan (2010, WSJ hereafter) considered a similar approach in order to find a few linear combinations of X that can capture most information of X in a nonlinear sense. The main difference between MIC and WSJ is the motivation and estimation of the components. WSJ is closer to the sufficient dimension reduction of Li (1991) in its motivation, while MIC is more in functional approximation.

The most *predictable* component of Hotelling (1935) is linear combination of variables in one set that can be best predicted by the variables in another set is called the most predictable component. Here MIC is the component that best predicts all the variables in the same set. For any MIC, $\beta_d^\top X$, consider the linear most predictable component, $\ell_d^\top X$, that minimizes

$$\|\ell^\top R_{d-1} - L(\ell^\top R_{d-1}|\beta_d^\top X)\|$$

with respect to ℓ with $\|\ell\| = 1$. If $\beta_d^\top X$ is a PC, then $\ell_d = \beta_d$, but if $\beta_d^\top X$ is a MIC, ℓ_d can differ from β_d. Let $\tilde{R}_{d-1} = R_{d-1} - L(R_{d-1}|\beta_d^\top X)$. We take the nonlinear most predictable component $\gamma_d^\top X$ to minimize

$$E\|\gamma^\top \tilde{R}_{d-1} - E(\gamma^\top \tilde{R}_{d-1}|\beta_d^\top X)\|^2$$

with respect to γ with $\|\gamma\| = 1$. Thus, $\beta_d^\top X$ can best predict $\gamma_d^\top X$ nonlinearly. See also Li (1997).

4. Estimation of the Most Informative Components

The estimation of MICs is related to the single-index model for which there are many efficient estimation methods. See for example Härdle, Hall, and Ichimura (1993), Härdle and Stoker (1989), Hristache, Juditski, and Spokoiny (2001), Yu and Ruppert (2002), Yin and Cook (2005), and Xia (2006). Because the problem here is "unsupervised", the main difficulty is to find an appropriate initial value in implementing the estimation; existing methods such as Härdle and Stoker

(1989) or the outer product of gradients method (Samarov (1993); Xia, Tong, and Li (2007)) cannot be used directly.

Based on the definition of MIC, we need to consecutively estimate β_d, $d = 1, \dots$, to minimize

$$E\|R_{d-1} - E(R_{d-1}|\beta_d^\top X)\|^2.$$

This minimization can be implemented as follows. First, employ a nonparametric smoothing regression method such as splines or kernel smoothing to estimate $g_\beta^{[k]}(u) = E(R_k|\beta^\top X = u)$ for any β. For sample $X_1, \dots, X_n$, let $R_{01} = X_1, \dots, R_{0n} = X_n$. Using the local linear kernel smoothing, $g_\beta(u)$ can be estimated by

$$\hat{g}_\beta^{[1]}(u) = \frac{\sum_{i=1}^n w_{n,i}^\beta(u) R_{0i}}{\sum_{i=1}^n w_{n,i}^\beta(u)}, \tag{4.1}$$

where $w_{n,i}^\beta(u) = s_n^{(2)} K_b(\beta^\top X_i - u) - s_n^{(1)} K_b(\beta^\top X_i - u)\{(\beta^\top X_i - u)/b\}$ and $s_n^{(k)} = n^{-1} \sum_{i=1}^n K_b(\beta^\top X_i - u)\{(\beta^\top X_i - u)/b\}^k$. Here $K(\cdot)$ is a kernel function, b is the bandwidth and $K_b(\cdot) = K(./b)/b$, see Fan and Gijbels (1996). The first MIC is the minimizer

$$\hat{\beta}_1 = \arg\min_{\beta:\|\beta\|=1} n^{-1} \sum_{j=1}^n \|R_{0j} - \hat{g}_\beta^{[1]}(\beta^\top X_j)\|^2. \tag{4.2}$$

Details for this minimization can be found in Xia (2007) where an iterative algorithm is provided with a closed form for each iteration.

After the first MIC is obtained, denoted by $\hat{\beta}_1$, we can calculate $\hat{g}_{\hat{\beta}_1}(\cdot)$ according to (4.1) with residuals

$$R_{1i} = R_{0i} - \hat{g}_{\hat{\beta}_1}^{[1]}(X_i), \quad i = 1, \dots, n.$$

With R_{1i} in (4.1), the minimizer of β is the second MIC, denoted by $\hat{\beta}_2$. Continuing this procedure, we can get the estimators for MICs, denoted by $\hat{\beta}_1, \hat{\beta}_2, \dots$.

When the bandwidth is sufficiently large, the local linear kernel estimator is the linear regression.

Theorem 1. *For any fixed sample, as bandwidth $b \to \infty$, $\hat{\beta}_d$ tends to the dth PC of the sample.*

We propose a method to obtain an appropriate initial estimator for this minimization. Consider the linear PCA again. Suppose the probability density function of X is $f(x)$ and write $\triangledown f(x) = \partial f(x)/\partial x$ for the gradient of $f(x)$. If $X \sim N(\mu, \Sigma)$, then we have the gradient of its probability density function as

$$\triangledown f(x) = -(2\pi)^{-p/2} Det(\Sigma)^{-1/2} \exp\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\}\Sigma^{-1}(x-\mu).$$

Thus $E\{\triangledown f(X)\triangledown^\top f(X)\} = c\Sigma^{-1}$, where $c = \sqrt{3}(2\pi)^{-p}Det(\Sigma)^{-1}$. Therefore, the PCs can be obtained by the eigenvectors of $E\{\triangledown f(X)\triangledown^\top f(X)\}$.

Lemma 1. *Suppose $EX = 0$ and $Cov(X) = \Sigma$, and that there are vectors $\theta_1, \theta_2, \ldots, \theta_p$ of full rank such that $\theta_k^\top X = g_k(\theta_1^\top X) + \varepsilon_k$, and that the joint distribution of $(\varepsilon_2, \ldots, \varepsilon_p)$ and $\theta_1^\top X$ is normal. If there is at least one g_k that is neither a linear function nor a constant, then θ_1 is an eigenvector of $E\{\triangledown f(X)\triangledown^\top f(X)\}$.*

The justification here is for the normal variables. The justification under a general framework needs investigation. Based on the lemma, we can obtain an initial estimator using the kernel density estimation as follows. Suppose $H(\cdot)$ is a p-dimensional kernel function and h is the bandwidth. Then the density function $f(x)$ can be estimated by

$$\hat{f}(x) = n^{-1}\sum_{i=1}^{n} H_h(X_i - x),$$

where $H_h(\cdot) = h^{-p}H(./h)$. The sample version of the gradient of the density function is

$$\triangledown\hat{f}(x) = n^{-1}\sum_{i=1}^{n} \triangledown H_h(X_i - x),$$

where $\triangledown H_h(X_i - x) = h^{-p-1}\partial H((X_i - x)/h)/\partial x$. We can estimate Σ by

$$S_n = n^{-1}\sum_{i=1}^{n} \triangledown\hat{f}(X_i)\triangledown^\top\hat{f}(X_i).$$

Lemma 2. *Suppose assumptions* (A1)−(A3) *in the appendix hold. If $h \to 0$, $nh^{p+1} \to \infty$, and a kth order kernel is used,*

$$\|S_n - E\{\triangledown f(X)\triangledown^\top f(X)\}\| = O_p(h^{k+1} + n^{-1/2}).$$

Denote the eigenvectors of S_n by $\hat{\theta}_1, \ldots, \hat{\theta}_p$, and consider the approximation errors

$$n^{-1}\sum_{j=1}^{n} \|R_{0j} - \hat{g}^{[1]}_{\hat{\theta}_k}(\hat{\theta}_k^\top X_j)\|^2, \quad k = 1, \ldots, p.$$

The eigenvector corresponding to the smallest approximation error is the direction used as the initial estimator $\hat{\beta}_1$. Under the assumptions of Lemma 1, this initial estimator has a decent convergence rate. When k is large, $\sqrt{n}h^{k+1} \to 0$, and other conditions for h are satisfied, $\hat{\beta}_1$ is root-n consistent. After the initial value is obtained, we refine the estimator by minimizing (4.2).

Remark 1. The asymptotic distribution of MIC is not easy to investigate. For one, the structure of the component is not clear, as shown in Example 2. We can work some special cases. For example, suppose there is a matrix B such that

$$B^\top X = (g_2(\theta_1^\top X), \ldots, g_p(\theta_1^\top X))^\top + (\varepsilon_2, \ldots, \varepsilon_p)^\top$$

and that (B, θ_1) is of full rank and orthogonal. If $\varepsilon_2, \ldots, \varepsilon_p$ and $\theta_1^\top X$ are independent and normally distributed, and if θ_1 is the first MIC then, following the proofs of Härdle, Hall, and Ichimura (1993) or Xia (2007), we can prove that

$$\sqrt{n}(\hat{\beta}_1 - \theta_1) \to N(0, W_0^+ W_1 W_0^+)$$

in distribution, where $W_0 = \sum_{k=2}^p E\{(g_k'(\theta_1^\top X))^2 (X - E(X|\theta_1^\top X))(X - E(X |\theta_1^\top X))^\top\}$ and $W_1 = \sum_{k=2}^p E\{(g_k'(\theta_1^\top X))^2 (X - E(X|\theta_1^\top X))(X - E(X|\theta_1^\top X))^\top\} Var(\varepsilon_k)$.

In this estimation, two bandwidths h and b are involved. Calculations suggest that a symmetric density function is stable as the kernel function, for which the commonly used bandwidth is $h = 2.34 n^{-1/(p+4)}$ when the data are standardized and the Epanechnikov kernel is used. See Scott (1992) for more details. Similarly, for bandwidth b, after standardizing all the variables, we use the rule-of-thumb bandwidth $h = 2.34 n^{-1/5}$. Of course, we can also use leave-one-out cross-validation to select the bandwidths; see Silverman (1986).

5. Identification of the Nonlinear Components

We first identify whether a MIC is indeed nonlinear or not.

Lemma 3. *A MIC, $\beta_d^\top X$, is linear if and only if*

$$E\|R_{d-1} - E(R_{d-1}|\beta_d^\top X)\|^2 = E\|R_{d-1} - L(R_{d-1}|\beta_d^\top X)\|^2. \qquad (5.1)$$

Based on this, we can identify linear MICs as follows. Suppose $\hat{\beta}_d$ is the estimator of β_d. Let $\tilde{R}_{d-1,i} = R_{d-1,i} - \hat{L}(R_{d-1,i}|\hat{\beta}_d^\top X_i)$, where $\hat{L}(R_{d-1,i}|\hat{\beta}_d^\top X_i) = \{\sum_{i=1}^n (\hat{\beta}_d^\top X_i)^2\}^{-1} \sum_{i=1}^n \hat{\beta}_d^\top X_i R_{d-1,i}$. The local linear leave-one-out estimate of $E(\tilde{R}_{d-1}|\hat{\beta}_d^\top X = u)$ is

$$\hat{V}_{-j}(u) = \frac{\sum_{i \neq j} w_{n,-j,i}(u) \tilde{R}_{d-1,i}}{\sum_{i \neq j} w_{n,-j,i}(u)},$$

where

$$w_{n,-j,i}(u) = s_{n,-j}^{(2)} K_b(\hat{\beta}_d^\top X_i - u) - s_{n,-j}^{(1)} K_b(\hat{\beta}_d^\top X_i - u) \Big\{ \frac{\hat{\beta}_d^\top X_i - u}{b} \Big\},$$

$$s_{n,-j}^{(k)} = n^{-1} \sum_{\ell \neq j} K_b(\hat{\beta}_d^\top X_\ell - u) \Big\{ \frac{\hat{\beta}_d^\top X_\ell - u}{b} \Big\}^k.$$

Let

$$CV(d) = n^{-1}\sum_{j=1}^{n}\|\tilde{R}_{d-1,j}\|^2 - n^{-1}\sum_{j=1}^{n}\|\tilde{R}_{d-1,j} - \hat{V}_{-j}(\hat{\beta}_d^\top X_j)\|^2.$$

If $\hat{\beta}_d^\top X$ is linear, $E(\tilde{R}_{d-1}|\hat{\beta}_d^\top X) = 0$. Thus, if $CV(d) \geq 0$, $\hat{\beta}_d^\top X$ is linear, otherwise it is nonlinear.

Following Theorem 5.2 of Xia (2007) we can show that this procedure is consistent in identifying whether the first MIC is linear or nonlinear, but consistency for the other MICs is complicated. When a MIC is identified as linear, we can make its estimator more efficient by taking a large bandwidth b.

6. Proportions of Variation Explained by the Components

After standardization, the total variation of $X = (\mathbf{x}_1, \ldots, \mathbf{x}_p)^\top$ is p. The variation of X explained linearly by the dth PC is its variance $\lambda_d = Var(\theta_d^\top X), d = 1, \ldots, p$. This is the variation explained in the following factor model

$$R_d = c + b\beta^\top X + \varepsilon,$$

where R_d is defined by (2.3), c and b are two vectors. The variation $Var(\theta_d^\top X)$ can be written as

$$\lambda_d = Var(\theta_d^\top X) = E\|L(R_d|\beta_{d+1}^\top X)\|^2,$$

and cumulative percentage of variation explained by the first d PCs is

$$C_L(d) = \frac{\lambda_1 + \cdots + \lambda_d}{\lambda_1 + \cdots + \lambda_p}.$$

It is common in practice to use a threshold, say 85%, to select the number of important PCs, and if $C_L(d) \geq 85\%$, then $\theta_1^\top X, \ldots, \theta_d^\top X$ contain the major information in X.

For the dth MIC we have

$$E\|R_{d-1}\|^2 = E\|E(R_{d-1}|\beta_d^\top X)\|^2 + E\|R_d\|^2.$$

In terms of variance analysis of regression, some of the variation in R_{d-1} is explained by MIC $\beta_d^\top X$, $E\|E(R_{d-1}|\beta_d^\top X)\|^2$. We take

$$c(d) = E\|E(R_{d-1}|\beta_d^\top X)\|^2$$

as the variation of R_{d-1} explained by $\beta_d^\top X$. The cumulative variation explained by $\beta_1^\top X, \ldots,\ \beta_d^\top X$ is then $c(1) + \cdots + c(d) = E\|X\|^2 - E\|R_d\|^2$. We take the cumulative variation of X explained by the first d MICs as

$$C_N(d) = \frac{c(1) + \cdots + c(d)}{\lambda_1 + \cdots + \lambda_p}.$$

Example 1. Consider $X = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)^\top$ with $\mathbf{x}_1, \mathbf{x}_2 \overset{IID}{\sim} N(0,1)$ and $\mathbf{x}_3 = \{(\mathbf{x}_1 + \mathbf{x}_2)^2 + c\varepsilon\}/\sqrt{8+c^2}$. It is easy to see that $\mathrm{Cov}(X) = diag(1,1,1)$. Thus, variables $\mathbf{x}_1, \mathbf{x}_2$, $\mathbf{x}_3$ are linearly uncorrelated and dimension cannot be reduced by PCA. Here, the first MIC is $F_1 = (\mathbf{x}_1 + \mathbf{x}_2)/\sqrt{2}$ and the second, $F_2 = (\mathbf{x}_1 - \mathbf{x}_2)/\sqrt{2}$. If $R_0 = X$, $R_1 = R_0 - E(R_0|F_1)$, and $R_2 = R_1 - E(R_1|F_2)$, then

$$E\|R_0\|^2 = 3, \quad E\|R_1\| = 1 + \frac{c^2}{8+c^2}, \quad E\|R_2\|^2 = \frac{c^2}{8+c^2}.$$

The total variation of X contained in F_1 and F_2 is $3 - c^2/(8+c^2) = E\|E(R_0|F_1)\|^2 + E\|E(R_1|F_2)\|^2$, while unexplained by F_1 and F_2 is $c^2/(8+c^2)$. Thus, two MICs are enough for data analysis when c is small.

Example 2. Nonlinear structure may not be of interest if it is not so strong as the linear dependence. Here is an example. Suppose ξ and ϵ are IID $N(0,1)$ and $X = (\mathbf{x}_1, \mathbf{x}_2)^\top = (\xi^2 - 1 + \xi + c\epsilon, \xi^2 - 1 - \xi + c\epsilon)^\top/\sqrt{3+c^2}$. We have

$$Cov(X) = \begin{pmatrix} 1, & (1+c^2)/(3+c^2) \\ (1+c^2)/(3+c^2), & 1 \end{pmatrix}.$$

Both PCs and MICs are $F_a = 2(\xi^2 - 1 + c\epsilon)/\sqrt{2(3+c^2)}$ and $F_b = 2\xi/\sqrt{2(3+c^2)}$ but with possibly different order. Component F_a is linear in X, but F_b is not. Under the PC framework, the information contained in F_a and F_b are

$$Var(F_a) = 2(2+c^2)/(3+c^2) \quad \text{and} \quad Var(F_b) = 2/(3+c^2).$$

Because $Var(F_a) > Var(F_b)$, F_a is the first PC and F_b the second.

The variation of X explained nonlinearly by F_b alone is $6/(3+c^2)$. If $c < 1$, then $2(2+c^2)/(3+c^2) < 6/(3+c^2)$ and thus F_b is the first MIC and is nonlinear; F_a is the second MIC and is used only to predict $R_1 = (c\epsilon, c\epsilon)^\top/\sqrt{3+c^2}$. However, if $c > 1$, then F_a is the first MIC and linear, while F_b's contribution is only $2/(3+c^2)$ after F_a is used. In that case, no nonlinear MIC is used.

7. Numerical Studies

We used simulated data to check the efficiency of the proposed estimation method and the identification method. We had four real data sets to illustrate the application of MIC in clustering, in understanding the data structure, and in dimension reduction of X for regression analysis. Comparison was also made between MIC and PC in the applications. All variables in the data were standardized separately before the methods are used.

Table 1. Simulation results for Example 3 with $p = 5$.

					nonlinearity of MICs		
p	n	$e(\hat{\beta}_1)$	$e(\hat{\beta}_2)$	$e(\hat{\beta}_3)$	1st	2nd	3rd
5	100	0.1750	0.3534	0.3286	1.00	0.99	0.12
	200	0.1648	0.2003	0.1630	1.00	1.00	0.08
	500	0.1169	0.1191	0.0951	1.00	1.00	0.06

Example 3. (simulations) Suppose $(\xi_1, \ldots, \xi_p)^\top \sim N(0, \Sigma)$ with $\Sigma = (0.5^{|i-j|})_{1 \le i,j \le p}$, and that $\mathbf{z}_1 = \xi_1, \mathbf{z}_2 = \xi_2, \mathbf{z}_3 = c_3(\cos(2\xi_1) + 0.2\varepsilon_1)$, $\mathbf{z}_4 = c_4(|\xi_5| + 0.2\varepsilon_2)$, $\mathbf{z}_k = \xi_k, k \ge 5$, where c_3 and c_4 are selected such that $Var(\mathbf{z}_3) = Var(\mathbf{z}_4) = 1$. Let $Z = (\mathbf{z}_1, \ldots, \mathbf{z}_p)^\top$. We observed

$$X = V_0 Z, \quad \text{with } V_0 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & -\frac{\sqrt{2}}{2} & 0 \\ -\frac{1}{2} & \frac{1}{2} & \frac{\sqrt{2}}{2} & 0 & 0 \\ -\frac{1}{2} & \frac{1}{2} & 0 & \frac{\sqrt{2}}{2} & 0 \\ -\frac{1}{2} & \frac{1}{2} & -\frac{\sqrt{2}}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & I_{p-4} \end{pmatrix},$$

where I_{p-4} is the identity matrix of $(p-4) \times (p-4)$. We took the estimation error of an estimator $\hat{\beta}_k$ for the true parameter β_k with $\|\beta_k\| = 1$ as

$$e(\hat{\beta}_k) = \sqrt{1 - (\hat{\beta}_k^\top \beta_k)^2}.$$

For $p = 5$, the first three MICs are $\beta_1^\top X, \beta_2^\top X$ and $\beta_3^\top X$ with $\beta_1 = (0.5, 0.5, 0, \sqrt{2}/2, 0, \ldots, 0)^\top$, $\beta_2 = (0, 0, 0, 0, 0, 1, 0, \ldots, 0)^\top$, and $\beta_3 = (-0.67,\ 0.23,\ 0.63,\ 0.31,\ -0.11)^\top$. The first two components are nonlinear and the third linear. These components together explain about 85% of variation of X. For $p = 10$, the first six MICs are $\beta_1 = (0,\ 0,0,0,0,\ 0.36,\ 0.48,\ 0.52,\ 0.48,\ 0.36)^\top$, $\beta_2 = (0,\ 0,\ 0,\ 0,\ 0,\ 1,\ 0,\ \ldots,0)^\top$, $\beta_3 = (0.5,\ 0.5,\ 0,\ \sqrt{2}/2,\ 0,\ \ldots,0)^\top$, $\beta_4 = (0,\ 0,0,0,0,\ \text{-}0.56\ \text{-}0.44,\ 0.00,\ 0.44,\ 0.56)^\top$, $\beta_5 = (-0.67,\ 0.23,\ 0.63,\ 0.31,\ \text{-}0.11,\ 0,0,0,0,0)^\top$, and $\beta_6 = (0,\ 0,\ 0,\ 0,\ 0,\ 0.56,\ \text{-}0.10,\ \text{-}0.59,\ \text{-}0.10,\ 0.56)^\top$. The second and third components are nonlinear and the others linear. These six components explain more than 85% of the variation of X. Based on 100 replications, the average estimation errors and the frequencies of identifying components as nonlinear are listed in Tables 1 and 2, respectively, for $p = 5$ and $p = 10$.

Tables 1 and 2 suggest that both our estimation and identification methods have satisfactory performance. As sample size n increases, the estimation errors decrease; the frequencies of identifying nonlinear components correctly tend to 1, and the frequencies of identifying linear components as nonlinear tend to 0.

With $p = 5$ the average CPU times were 18, 50 and 170 seconds, respectively, for $n = 100, 200$ and 500; with $p = 10$, the corresponding time were 47, 138 and

Table 2. Simulation results for Example 3 with $p = 10$.

							nonlinearity of MICs					
n	$e(\hat{\beta}_1)$	$e(\hat{\beta}_2)$	$e(\hat{\beta}_3)$	$e(\hat{\beta}_4)$	$e(\hat{\beta}_5)$	$e(\hat{\beta}_6)$	1st	2nd	3rd	4th	5th	6th
100	0.30	0.85	0.90	0.82	0.92	0.82	0.00	0.25	0.15	0.06	0.01	0.00
200	0.20	0.31	0.43	0.35	0.57	0.53	0.00	0.92	0.84	0.02	0.00	0.00
500	0.14	0.14	0.19	0.16	0.39	0.42	0.00	1.00	1.00	0.00	0.00	0.00

504 seconds, respectively, when the Intel Quad Q9650 3.0GHz processor was used.

Example 4 (Clustering)**.** In PCA, it is common to use the scatter plots of the first few PCs to cluster samples. MICs can be used for the same purpose. Since MIC is more efficient in detecting nonlinear patterns, it might be also more powerful in clustering complicated data sets. We considered data provided by Cook and Swayne (1995) for clustering that are thought difficult to cluster by the data providers (`http://www.ggobi.org/book/`).

Applying PCA, the variation explained by the PCs are 1.31, 1.24, 1.01, 0.98, and 0.46. It seems that there is no principal component that contributes a dominant portion of the variation. If we use the first two PCs, the scatter plot is shown in the first panel of Figure 1; the data are not clearly clustered. When we apply MIC, the variation explained by the MICs are 1.99, 1.00, 0.96, 0.74 and 0.15. The first MIC, with $\beta_1 = (0.03, -0.02, 0.72, -0.69, 0.03)^\top$, explains a good deal more than the others. If we plot the first MIC against its most predictable direction $\ell_1 = (0.07, -0.02, -0.70, -0.70, 0.15)^\top$, we obtain the second panel in Figure 1; it shows a nonlinear structure in the data, and that there are three clusters. By removing the linear part of the most predictable direction, we obtain panel 3 of Figure 1; the data are clearly separated into three clusters labeled A, B, and C.

Professor Dianne Cook, who provided the data, kindly pointed out that the three groups we clustered are correct but that there is another group hidden in one of them, suggesting that a hierarchical cluster approach is needed. We applied the same method to the three groups separately and found that group A can be further clustered into 2 subgroups labeled A_1 and A_2 in the last panel of Figure 1, while groups B and C cannot be further clustered.

Example 5 (Cars data)**.** These data were used by the American Statistical Association in its second exposition of statistical graphics technology in 1983. The data set is available at http://lib.stat.cmu.edu/datasets/cars.data. There are 406 observations on 8 variables: miles per gallon (X_1), number of cylinders (X_2), engine displacement (X_3), horsepower (X_4), vehicle weight (X_5), time to accelerate from

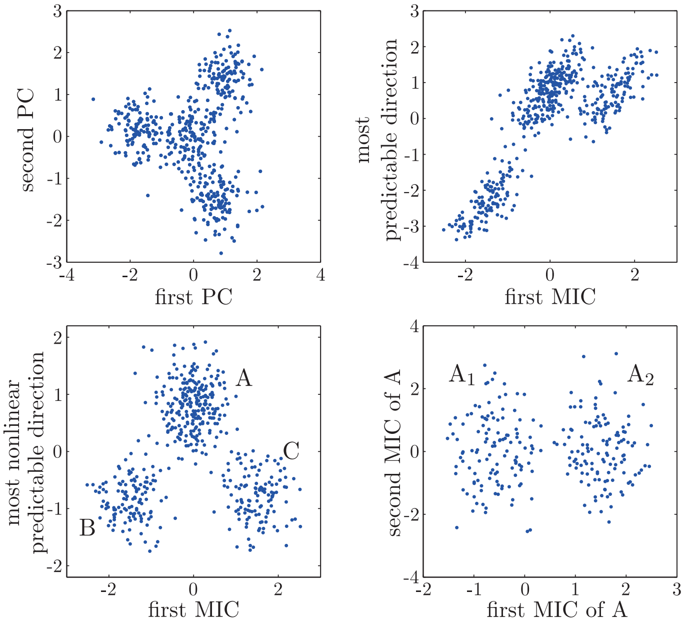

Figure 1. Clustering results for the data provided by Cook and Swayne (1995). The first panel is the scatter plot of the first PC against the second PC. The second panel is the plot of the first MIC against its most predicable direction. After removing the linear part in Panel 2, the third panel is obtained.

0 to 60 mph (X_6), model year (X_7), and origin of a car ((X_8, X_9): (1,0) indicates American, (0, 1) European, and (0,0) Japanese).

We first carry out the PCA analysis. The eigenvalues are 5.55, 1.28, 0.77, 0.69, 0.31, 0.18, 0.11, 0.05, and 0.03, and their cumulative contributions are 61.82%, 76.09%, 84.70%, 92.33%, 95.82%, 97.84%, 99.07%, 99.66%, and 100%. Applying MICA, the contribution of the components are 7.36, 0.59, 0.48, 0.16, 0.15, 0.08, 0.04, 0.03, and 0.03, and the cumulative contributions of the MICs are 81.80%, 88.34%, 93.65%, 95.46%, 97.08%, 97.91%, 98.39%, 98.73%, and 99.08%.

A possible reason for the first MIC to make such a large difference is as follows. The first MIC is nonlinear as shown in the first panel of Figure 2. There is a common linear dependence among variables of cars from the same origin but with shift differences betweens cars from different origins. As a comparison, we also plotted the first 2 PCs as shown in the panel on the right of Figure 2. The three clusters can also be identified, but not so clearly as with MIC.

Example 6. Another application of PCA is to linear regression when the covariates have collinearity, which is also called the principal component regression. MIC can also be used in nonparametric regression when the covariates have strong

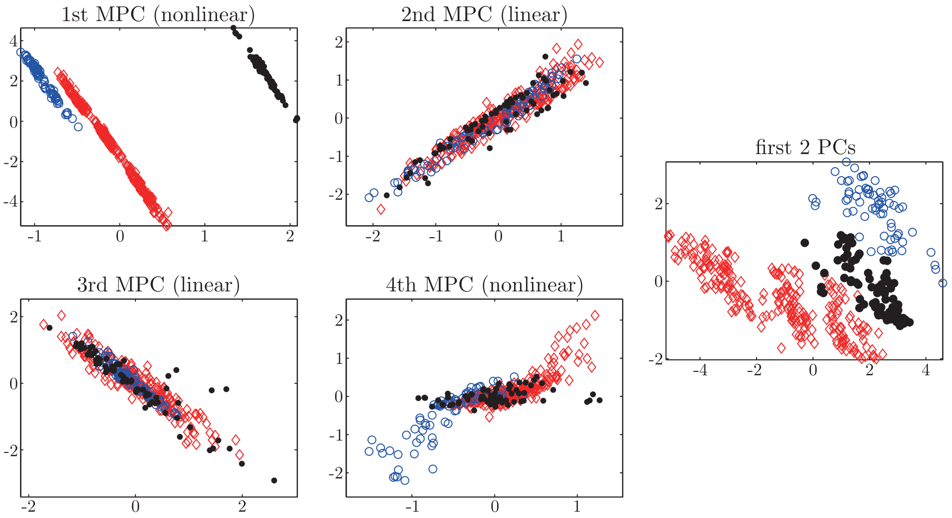

Figure 2. The four panels on the left hand side show the MIC analysis of the cars data. The panel on the right shows the plot of the first two PCs. In each panel, '•', '∘' and '⋄' represent cars from the USA, Europe and Japan respectively.

functional dependence. We applied MICA to the Boston Housing data that has been analyzed by Harrison and Daniel (1978), Doksum and Samarov (1995), and Fan and Huang (2005); the data are available at `http://cran.r-project.org/`. For each house, 13 variables were measured: $\mathbf{x}_1$: per capita crime rate by town; $\mathbf{x}_2$: proportion of residential land zoned for lots over 25,000 square feet; $\mathbf{x}_3$: proportion of non-retail business acres per town; $\mathbf{x}_4$: Charles River dummy variable, 1 if tract bounds river and 0 otherwise; $\mathbf{x}_5$: nitric oxides concentration in parts per 10 million; $\mathbf{x}_6$: average number of rooms per dwelling; $\mathbf{x}_7$: proportion of owner-occupied units built prior to 1940; $\mathbf{x}_8$: weighted distances to five Boston employment centers; $\mathbf{x}_9$: index of accessibility to radial highways; $\mathbf{x}_{10}$: full-value property-tax rate per \$10,000; $\mathbf{x}_{11}$: pupil-teacher ratio by town; $\mathbf{x}_{12}$: $(1{,}000(Bk-0.63)^2$ where Bk is the proportion of blacks by town; $\mathbf{x}_{13}$: percentage of lower status of the population.

We applied PCA and MICA to the data. The variations explained by the PCs and MICs are listed in Table 3. The first two MICs have obvious nonlinear contribution to X.Figure 3 further shows how the first MIC explains the variables nonlinearly.

To illustrate how MICs can help in nonparametric regression, consider a nonparametric model for the median value of owner-occupied homes in \$1,000's,

$$Y = g_D(C_1, \ldots, C_D) + \varepsilon, \tag{7.1}$$

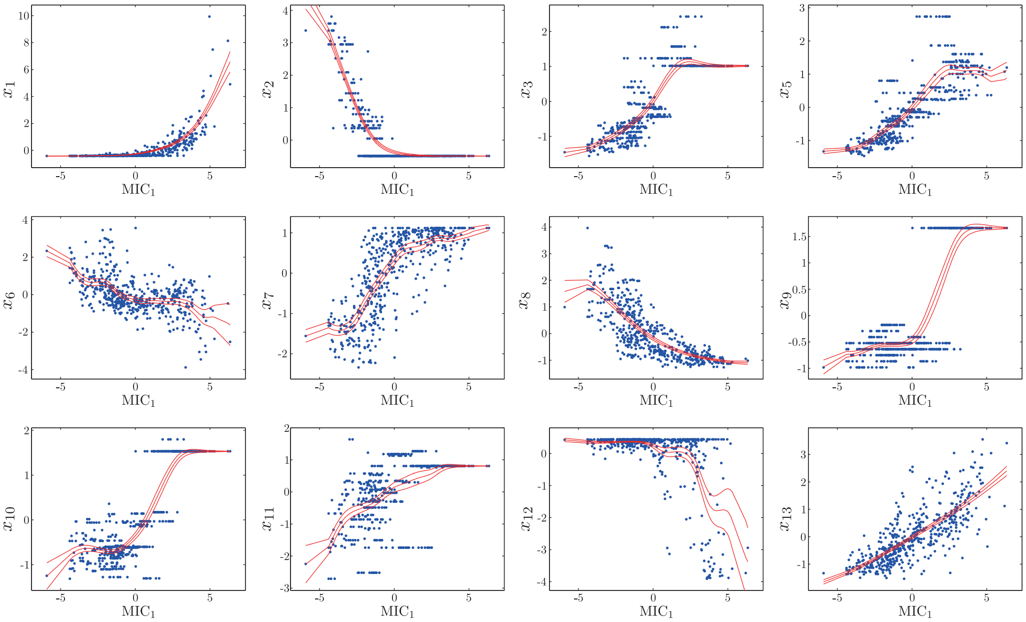

Figure 3. The nonlinear structure of each variable against the first MIC.

Table 3. The variation and comulative variantion explained by the components.

	PC		MIC			PC		MIC	
comp.	var.	cum.	var.	cum.	comp.	var.	cum.	var.	cum.
1	6.13	47.13%	7.70	59.23%	8	0.40	92.95%	0.11	96.69%
2	1.43	58.15%	2.67	79.77%	9	0.27	95.08%	0.07	97.23%
3	1.24	67.71%	0.74	85.46%	10	0.22	96.78%	0.05	97.63%
4	0.86	74.31%	0.42	88.69%	11	0.19	98.21%	0.04	97.92%
5	0.83	80.73%	0.39	91.69%	12	0.17	98.51%	0.03	98.15%
6	0.65	85.79%	0.37	94.54%	13	0.06	100%	0.02	98.31%
7	0.54	89.91%	0.17	95.85%					

where $C_k, k = 1, \ldots$ are the PCs or MICs with $D = 1, \ldots$. To compare the prediction ability of the model with different numbers of components, we partitioned the data into training set and testing set in the ratio of sample sizes 2:1 or 1:1. Based on the training set, we estimated the model using the k-nearest neighbor method with number of neighbors $k = 10$. We then applied the estimated model to predict the testing set. The prediction error was the mean of absolute differences between the true responses and predicted values. With 1,000 random partitions, the average of prediction errors are shown in Figure 4. With choices of k from 5 to 20, all the prediction errors had similar patterns.

Figure 4 has the model based on PCs improving prediction ability, but MICs

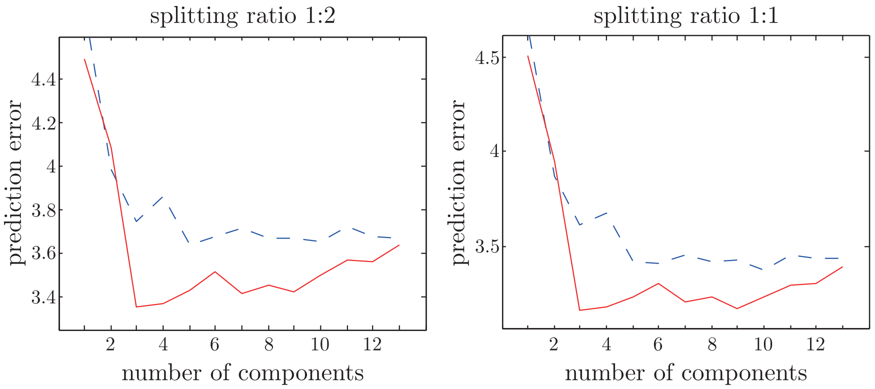

Figure 4. Prediction errors of the models based on the different ratio of training sets to testing sets. In each panel, the dashed line is the prediction error of model (7.1) based on the PCs, the solid line is that based on the MICs.

doing even better, and achieving the smallest error when three MICs are used. This example suggests that when there is functional structure in the variables, MIC is more powerful than PC in reducing the dimension of variables. Still better prediction for nonparametric regression may not always occur, its main contribution is to make model estimation more stable.

8. Conclusion

This paper has extended PCA to a more general framework in order to make it applicable to nonlinear structures in the variables. MICA is useful for unsupervised dimension reduction, and for extracting nonlinear features from X for better analysis of the data, such as clustering and pattern recognition. Some properties have been investigated.

Many problems of MICA need further investigation. For example, when nonlinearity can and should be detected. The asymptotic theory under general distribution assumptions requires work. Extension of MICA to time series data and "big p small n" problems are other important areas.

Acknowledgements

The authors thank an AE and two referees for their thoughtful comments. The research is partially supported by the Education Department of Nature Science Research of Guizhou Province (Grant No. 2010028), the Nomarch Foundation of Guizhou Province (Grant 2010025), China, the National Natural Science

Foundation of China (Grant 71371095), and NUS grant: R-155-000-121-112, Singapore.

Appendix: Assumptions and Proofs

Proof of Proposition 1. For any $p \times d$ matrix B, the best linear prediction of X based on B, $b + C^\top(B^\top X)$ minimizes $\min_{\beta,C} E\|X - b - C^\top B^\top X\|^2$. It is easy to see that if $EX = 0$ then $b = 0$ and $C = \{B^\top E(XX^\top)B\}^{-1}E(B^\top XX^\top)$. Note that

$$\begin{aligned} & E\|X - E(XX^\top)B\{B^\top E(XX^\top)B\}^{-1}B^\top X\|^2 \\ &\quad = E\|[I - E(XX^\top)B\{B^\top E(XX^\top)B\}^{-1}B^\top]X\|^2 \\ &\quad = tr[\{I - \Sigma B(B^\top \Sigma B)^{-1}B^\top\}\Sigma\{I - \Sigma B(B^\top \Sigma B)^{-1}B^\top\}^\top] \\ &\quad = tr[\Sigma - \Sigma B(B^\top \Sigma B)^{-1}B^\top \Sigma] \\ &\quad = \lambda_1 + \cdots + \lambda_p - tr[B^\top \Sigma^2 B\{B^\top \Sigma B\}^{-1}]. \end{aligned}$$

Let $D = \Sigma^{1/2}B$ and $\Gamma = D(D^\top D)^{-1/2}$. Then $\Gamma^\top\Gamma = I_d$ and

$$\begin{aligned} tr[B^\top \Sigma^2 B\{B^\top \Sigma B\}^{-1}] &= tr[D^\top \Sigma D\{D^\top D\}^{-1}] = tr[D^\top \Sigma D\{D^\top D\}^{-1}] = tr[\Gamma^\top \Sigma \Gamma] \\ &\le \lambda_1 + \cdots + \lambda_d. \end{aligned}$$

The last equality holds only when Γ is the first d eigenvectors of Σ. As a consequence, the best predictors are $B^\top X$ with $B = \Sigma^{-1/2}\Gamma$, which is the same base as Γ since $\Sigma^{-1/2}\Gamma = diag(\lambda_1^{-1/2}, \ldots, \lambda_d^{-1/2})\Gamma$.

Proof of Proposition 2. By Proposition 1 and letting $d = 1$, we have $\beta_1 = \theta_1$ and $R_1 = (I - \theta_1\theta_1^\top)X$. By induction, suppose for $1 \le d < p$, we have

$$\beta_1 = \theta_1, \ \ldots, \ \beta_d = \theta_d, \tag{A.1}$$

then $R_d = R_{d-1} - L(R_{d-1}|\theta_d^\top X) = \{I - \theta_d\theta_d^\top\}R_{d-1} = \{I - B_d B_d^\top\}X$, where $B_d = (\theta_1, \ldots, \theta_d)$. Let $\tilde{B}_d = (\theta_{d+1}, \ldots, \theta_p)$ and $\tilde{\beta} = (B_d, \tilde{B}_d)\beta$. It follows from $B_d^\top R_d = 0$ that

$$E\|R_d - L(R_d|\beta^\top X)\|^2 = E\|\tilde{B}_d^\top X - L(\tilde{B}_d^\top X|\tilde{\beta}^\top(B_d, \tilde{B}_d)^\top X)\|^2.$$

Note that $\tilde{B}_d^\top X$ is perpendicular to $B_d^\top X$. Thus,

$$L(\tilde{B}_d^\top X|\tilde{\beta}^\top(B_d, \tilde{B}_d)^\top X) = L(\tilde{B}_d^\top X|\tilde{\beta}_{(d)}\tilde{B}_d^\top X),$$

where $\tilde{\beta}_{(d)}$ is the last $p - d$ elements of $\tilde{\beta}$. If $\tilde{X} = \tilde{B}_d^\top X$, then

$$E\|R_d - L(R_d|\beta^\top X)\|^2 = E\|\tilde{X} - L(\tilde{X}|\tilde{\beta}_{d+1}^\top \tilde{X})\|^2.$$

By Proposition 1, $\tilde{\beta}_{d+1} = (1, 0, \ldots, 0)^\top$ minimizes $E\|\tilde{X} - L(\tilde{X}|\tilde{\beta}_d^\top \tilde{X})\|^2$. Thus $\beta_{d+1} = \tilde{B}_d\tilde{\beta}_{d+1} = \theta_{d+1}$, so (A.1) holds for $d + 1$. We thus complete the proof.

Proof of Proposition 3. This follows immediately from Proposition 2.

Proof of Lemma 1. Let $\Sigma_1 = Cov[(\varepsilon_2,\dots,\varepsilon_p)^\top]$ and take $(\theta_2,\dots,\theta_p)^\top := \Sigma_1^{-1/2}(\theta_2,\dots,\theta_p)^\top$, $(g_2,\dots,g_p)^\top := \Sigma_1^{-1/2}(g_2,\dots,g_p)^\top$ and $(\varepsilon_2,\dots,\varepsilon_p)^\top := \Sigma_1^{-1/2}$ $(\varepsilon_2,\dots,\varepsilon_p)^\top$. Then $\theta_k^\top X = g_k(\theta_1^\top X) + \varepsilon_k$, where $(\varepsilon_2,\dots,\varepsilon_p)^\top \sim N(0, I_{p-1})$. Define $\tilde X = \Sigma^{-1/2}X, \tilde\theta_k = \Sigma^{1/2}\theta_k/c_k$, $\tilde\varepsilon_k = \varepsilon_k/c_k$, where $c_k = \|\Sigma^{1/2}\theta_k\| k = 2,\dots,p$, and $\tilde\theta_1 = \theta_1/c_1$ with $c_1 = var(\theta_1^\top X)$ and $\tilde g_k(u) = g_k(c_1u)/c_2$. Then we have $E\tilde X = 0$, $Cov(\tilde X) = I$, $Var(\tilde\theta_k^\top \tilde X) = 1$, $\theta_k^\top \tilde X = \tilde g_k(\tilde\theta_1^\top \tilde X) + \tilde\varepsilon_k$, and that $(\tilde\varepsilon_2,\dots,\tilde\varepsilon_p)$ and $\tilde\theta_1^\top \tilde X$ are independent. Without loss of generality, we take

$$\tilde\theta_1^\top \tilde\theta_k = 0, k = 2,\dots,p. \tag{A.2}$$

Otherwise, we redefine $\tilde g_k(\theta_1^\top X) := \tilde g_k(\tilde\theta_1^\top X) - \tilde\theta_1^\top X \tilde\theta_1^\top \tilde\theta_k$, $\tilde\theta_k := \tilde\theta_k - \tilde\theta_1\tilde\theta_1^\top\tilde\theta_k$ and $\tilde\theta_k := \tilde\theta_k/(\tilde\theta_k^\top\tilde\theta_k)^{1/2}$.

Let $Z = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_p)^\top = (\tilde\theta_1, \tilde\theta_2, \dots, \tilde\theta_p)^\top \tilde X$. By assumption, we have $E\{Z\} = 0$ and $Cov(Z) = diag(1,\dots,1)$. Thus, with $\sigma_k^2 = Var(\tilde\varepsilon_k)$,

$$E\{\mathbf{z}_1(\mathbf{z}_2,\dots,\mathbf{z}_p)^\top\} = 0, \quad Var(\mathbf{z}_1) = 1,$$

$$Var(\mathbf{z}_k) = E[\{\tilde g_k(\mathbf{z}_1)\}^2] + \sigma_k^2 = 1, \quad k = 2,\dots,p.$$

Because $cov(\mathbf{z}_1, \mathbf{z}_k) = 0$, it follows that $E\{\mathbf{z}_1\tilde g_k(\mathbf{z}_1)\} = 0$, $k = 2,\dots,p$. The density function of $\mathbf{z}_1$ is $f_\xi(v) = (2\pi)^{-1/2}\exp\{-v^2/2\}$, so $f_\xi'(v) = -vf_\xi(v)$ and

$$\begin{aligned} E\{\tilde g_k'(\mathbf{z}_1)\} &= \int_{-\infty}^{\infty} \tilde g_k'(v) f_\xi(v)dv = -\int_{-\infty}^{\infty} f_\xi'(v)g(v)dv \\ &= \int_{-\infty}^{\infty} v\tilde g_k(v) f_\xi(v)dv = E\{\xi\tilde g_k(\xi)\} = 0. \end{aligned} \tag{A.3}$$

The joint probability density function of Z is

$$f_Z(z_1,\dots,z_p) = \{2\pi\}^{-p/2}(\prod_{k=2}^p \sigma_k)^{-1}\exp\{-\frac{z_1^2}{2} - \sum_{k=2}^p \frac{(z_k - \tilde g_k(z_1))^2}{2\sigma_k^2}\}.$$

Therefore, we have

$$\frac{\partial f_Z(z)}{\partial z} = f_Z(z_1,\dots,z_p)\Psi,$$

where

$$\Psi = \{-z_1 + \sum_{k=2}^p \frac{(z_k - \tilde g_k(z_1))\tilde g_k'(z_1)}{\sigma_k^2}, -\frac{(z_2 - \tilde g_2(z_1))}{\sigma_2^2}, \dots, -\frac{(z_p - \tilde g_p(z_1))}{\sigma_p^2}\}^\top.$$

Let

$$\Lambda_0 = E\Big\{\frac{\partial f_Z(Z)}{\partial z}\Big(\frac{\partial f_Z(Z)}{\partial z}\Big)^\top\Big\} = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f_Z^3(z_1,\dots,z_p)\Psi\Psi^\top dz_1\cdots dz_p.$$

Based on the assumptions and (A.3), calculation shows

$$\Lambda_0 = \frac{1}{3}\sqrt{3}^p(2\pi)^{-p}(\prod_{k=2}^{p}\sigma_k)^{-2}diag\Big(1+\sum_{k=2}^{p}\frac{E[\{\tilde{g}_k'(\sqrt{3}\xi)\}^2]}{\sigma_k^2}, \frac{1}{\sigma_2^2}, \dots, \frac{1}{\sigma_p^2}\Big).$$

Because $Z = (\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_p)^\top \tilde{X}$ with $(\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_p)^\top(\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_p) = I_p$, we have

$$\begin{aligned} E\{\triangledown f(\tilde{X})\triangledown^\top f(\tilde{X})\} &= (\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_p)E\Big\{\frac{\partial f_Z(Z)}{\partial z}\Big(\frac{\partial f_Z(Z)}{\partial z}\Big)^\top\Big\}(\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_p)^\top \\ &= (\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_p)\Lambda_0(\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_p)^\top. \end{aligned}$$

Therefore, $\tilde{\theta}_1$ is one of the eigenvectors of $E\{\triangledown f(\tilde{X})\triangledown^\top f(\tilde{X})\}$, and $\theta_1 = c_1\tilde{\theta}_1$.

To prove Lemma 1, we need assumptions.

(A1) Random variable X is bounded.

(A2) The density function of X has a $(d+2)$th order bounded derivative.

(A3) The kernel function $H(x)$ has bounded support, satisfies $\int H(v)dv_1 \cdots dv_p = 1$,

$$\int v_1^{d_1}\cdots v_p^{d_p}H(x) = 0, \text{ for any } d_1 + \cdots + d_p \le k,$$

and $\int \|u\|^{2k}H(u)du < \infty$.

Proof of Lemma 1. Write the estimator as

$$\triangledown \hat{f}(x) = \triangledown f(x) + M(x)h^{k+1} + \Delta_n(x) + \delta_n(x),$$

where

$$\Delta_n = \frac{1}{nh^{p+1}}\sum_{i=1}^{n}\Big(\triangledown H\Big\{\frac{X_i - x}{h}\Big\} - E\Big[\triangledown H\{\frac{X_i - x}{h}\}\Big]\Big),$$

with $\triangledown H(v) = \partial H(v)/\partial v$, and $\delta_n(x) = E\triangledown\hat{f}(x) - \triangledown f(x) - M(x)h^{k+1}$. By Masry (1996), we have $\delta_n(x) = o_p(h^{k+1})$. It follows that

$$\begin{aligned} &n^{-1}\sum_{j=1}^{n}\triangledown\hat{f}(X_j)\triangledown^\top\hat{f}(X_j) \\ &= n^{-1}\sum_{j=1}^{n}\triangledown f(X_j)\triangledown^\top f(X_j) + n^{-1}\sum_{j=1}^{n}\{M(X_j)\triangledown^\top f(X_j) + \triangledown f(X_j)M^\top(X_j)\}h^{d+1} \\ &\quad + n^{-1}\sum_{j=1}^{n}\{\Delta_n(X_j)\triangledown^\top f(X_j) + \triangledown f(X_j)\Delta_n^\top(X_j)\} + o_p(n^{-1/2} + h^{k+1}). \end{aligned}$$

We have

$$Var(n^{-1}\sum_{j=1}^{n}\{\Delta_n(X_j)\triangledown^\top f(X_j)+\triangledown f(X_j)\Delta_n^\top(X_j)\})=O(n^{-1}),$$

and the third term here is of $O_p(n^{-1/2})$.

Proof of Lemma 3. If $\beta_d^\top X$ is linear, then (5.1) holds. Let $\tilde{R}_{d-1}=R_{d-1}-L(R_{d-1}|\beta_d^\top X)$. It is easy to see that $E(\tilde{R}_{d-1}|\beta_d^\top X)=E(R_{d-1}|\beta_d^\top X)-L(R_{d-1}|\beta_d^\top X)$. With $\tilde{R}_{d-1}=E(\tilde{R}_{d-1}|\beta_d^\top X)+\{\tilde{R}_{d-1}-E(\tilde{R}_{d-1}|\beta_d^\top X)\}$,

$$E\|\tilde{R}_{d-1}\|^2=E\|E(\tilde{R}_{d-1}|\beta_d^\top X)\|^2+E\|\{\tilde{R}_{d-1}-E(\tilde{R}_{d-1}|\beta_d^\top X)\}\|^2.$$

By (5.1), we have $E\|E(R_{d-1}|\beta_d^\top X)-L(R_{d-1}|\beta_d^\top X)\|^2=0$. Thus, $E(R_{d-1}|\beta_d^\top X)=L(R_{d-1}|\beta_d^\top X)$.

References

Cheng, B. and Tong, H. (1992). On consistent nonparametric order determination and chaos. *J. Roy. Statist. Soc. Ser. B* **54**,427-449.

Comon, P. (1994). Independent component analysis: a new concept? *Signal Processing* **36**, 287-314

Cook, R. D. (2007). Fisher Lecture: Dimension Reduction in Regression. *Statist. Sci.* **22**, 1-26.

Cook, R. D. (2008). *Regression Graphics: Ideas for Studying Regressions Through Graphics*, Wiley, New York.

Cook, D. and Swayne, D. F. (1995). *Interactive and Dynamic Graphics for Data Analysis: With Examples Using R and GGobi.* Springer, New York.

Doksum, K. and Samarov, A. (1995) Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. *Ann. Statist.* **23**, 1443-1864

Fan, J. and Huang, T. (2005) Profile Likelihood Inferences on semiparametric varying-coefficient partially linear models. *Bernoulli* **11**, 1031-1057.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman & Hall, London.

Girard, S. and Iovleff, S. (2001). Auto-associative models, nonlinear Principal component analysis, manifolds and projection pursuit. Manuscript.

Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21**, 157-178

Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84**, 986-995.

Harrison, D. and Daniel, L. R. (1978). Hedonic housing prices and the demand for clean air. *J. Environmental Economics and Management* **5**, 81-102.

Hastie, T. and Stuetzle, W. (1989). Principal curves. *J. Amer. Statist. Assoc.* **84**, 502-516.

Hotelling, H. (1935). The most predictable criterion. *J. Educational Psychology* **26**, 139-142.

Hristache, M., Juditski, A. and Spokoiny, V. (2001). Direct estimation of the index coefficients in a single-index model. *Ann. Statist.* **29**, 595-623.

Huber, P. J. (1985). Projection pursuit. *Ann. Statist.* **13**, 435-475.

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE J.* **37**, 233-243.

Li, K. C. (1991) Sliced Inverse Regression for Dimension Reduction. *J. Amer. Statist. Assoc.*, **86**, 316-327.

Li, K. C. (1997). Nonlinear confounding in high-dimensional regression. *Ann. Statist.* **25**, 577-612.

Masry, E. (1996). Multivariate local polynomial regression for time series: uniform strong consistency and rates. **17**, 571-599.

Pan, J. and Yao, Q. (2008). Modelling multiple time series via common factors. *Biometrika* **95**, 365-379.

Park, B., Mammen, E., Härdle, W. and S. Borak (2009). Time series modelling with semiparametric factor dynamics. *J. Amer. Statist. Assoc.* **104**, 284-298.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **2**, 559-572.

Samarov, A. M. (1993). Exploring regression structure using nonparametric functional estimation. *J. Amer. Statist. Assoc.* **88**, 836-847.

Schölkopf, B., Smola, A. J. and Müller, K. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation* **10**, 1299-1319.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization.* Wiley, New York.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Wang, M., Sha, F. and Jordan, M. I. (2010). Unsupervised kernel dimension reduction. Proceedings of Neural Information Processing Systems. Vancouver, CA.

Xia, Y. (2006). Asymptotic distributions for two estimators of the single-index model. *Econometric Theory* **22**, 1112-1137.

Xia, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *Ann. Statist.* **35**, 2654-2690.

Xia, Y., Tong, H. and Li, W. K. (2002) Single-index volatility models and estimation. *Statist. Sinica* **12**, 785-799.

Yin, X. and Cook, R. D. (2005). Direction estimation in single-index regressions. *Biometrika* **92**, 371-384.

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single index models. *J. Amer. Statist. Assoc.* **97**, 1042-1054.

School of Management and Economics, University of Electronic Science and Technology of China. Chengdu, Sichuan, 610051, China.

E-mail: jingyaping@mail.gzife.edu.cn

School of Mathematics and Statistics, Guizhou University of Finance and Economics, Guiyang, Guizhou Province, 550025, P.R. China.

E-mail: honglei77.zhao@gmail.com

Department of Statistics and Applied Probability, National University of Singapore.

E-mail: staxyc@stat.nus.edu.sg

(Received September 2012; accepted July 2013)