Communications for Statistical Applications and Methods
2017, Vol. 24, No. 1, 97–113

http://dx.doi.org/10.5351/CSAM.2017.24.1.097
Print ISSN 2287-7843 / Online ISSN 2383-4757

Model-based inverse regression for mixture data

Changhwan Choi[a], Chongsun Park[1,a]

[a]Department of Statistics, Sungkyunkwan University, Korea

Abstract

This paper proposes a method for sufficient dimension reduction (SDR) of mixture data. We consider mixture data containing more than one component that have distinct central subspaces. We adopt an approach of a model-based sliced inverse regression (MSIR) to the mixture data in a simple and intuitive manner. We employed mixture probabilistic principal component analysis (MPPCA) to estimate each central subspaces and cluster the data points. The results from simulation studies and a real data set show that our method is satisfactory to catch appropriate central spaces and is also robust regardless of the number of slices chosen. Discussions about root selection, estimation accuracy, and classification with initial value issues of MPPCA and its related simulation results are also provided.

Keywords: dimension reduction, sliced inverse regression, mixture modeling, principal component analysis, probability model

1. Introduction

Consider a regression problem with response variable y and p predictors $\boldsymbol{x} = (x_1, \dots, x_p)^T$. The main interest of sufficient dimension reduction (SDR) method is retrieving $d < p$ linear combinations of $\boldsymbol{x}$ containing all structural information about y. If we know them, it is sufficient to investigate the $d + 1$ dimensional subspace instead of $p + 1$ dimensional space. Powerful sufficient dimension reduction methods have been proposed such as sliced inverse regression (SIR) (Li, 1991), sliced average variance estimation (SAVE) (Cook and Weisberg, 1991), principal Hessian direction (PHD) (Li, 1992), and directional regression (DR) (Li *et al.*, 2005). They all use sample version of the first or second moment of $\boldsymbol{x}|y$ in order to estimate the d linear combinations of $\boldsymbol{x}$. Scrucca (2011) proposed model-based SIR (MSIR) using a multiple local means fitted by a finite Gaussian mixture model (GMM) to obtain information about $\boldsymbol{x}|y$ to show that the method is more flexible to accommodate various regression functions including symmetric data. However, in the presence of mixture groups in data, existing SDR methods can mislead the results because they treat all observations as coming from one group.

In this paper, we introduce an SDR method applying the GMM approach of MSIR based on the mixture probabilistic principal component analysis (MPPCA) (Tipping and Bishop, 1999b) for mixture data. The SIR problem can be expressed as a generalized eigenvalue decomposition problem for original predictors and becomes a usual eigenvalue decomposition problem with standardized predictors. Tipping and Bishop (1999a, 1999b) introduced a latent variable model for the eigenvalue decomposition problem of principal component analysis that enables us to use and utilize a likelihood idea when developing a unified MSIR algorithm for mixture data. Under the assumption that y was

[1] Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2 Sungkyunkwan-ro, Jongno-gu, Seoul 03063, Korea. E-mail: cspark@skku.edu

Published 31 January 2017 / journal homepage: http://csam.or.kr

drawn from M groups, and each of them have distinct central subspaces (that is, none of the central subspaces is included in or equivalent to those of other groups), we could therefore obtain information on the $\boldsymbol{x}|y$ of each group by feeding multiple local means from GMM into a MPPCA model.

This paper is organized as follows. In Section 2, we briefly review SIR and MSIR with a definition of SDR. Probabilistic principal component analysis (PPCA) and its mixture model, MPPCA is described in the next section. We develop an SDR strategy for mixture data and discuss some of its problems in implementation in Section 4. Section 5 includes various simulation studies with results. A simple real data example is in Section 6. Some discussion is provided in Section 7.

2. SIR and model-based SIR

Cook (1998) defined SDR subspace as a column space of $B \in \mathbb{R}^{p\times d},\ d < p$ satisfying that

$$y \amalg \boldsymbol{x}|B^T\boldsymbol{x}, \tag{2.1}$$

where $\amalg$ stands for independence and central subspace $S_{y|\boldsymbol{x}}$ as all intersections of SDR which is not unique. Now, consider $z = \Sigma^{-1/2}[\boldsymbol{x} - E(\boldsymbol{x})]$, where $\Sigma = Cov(\boldsymbol{x})$ with corresponding central subspace of $y|z$ denoted by $S_{y|z}$, and let $\eta \in \mathbb{R}^{p\times d}$ be its orthonormal basis. Then, if A is a full rank p by p matrix

$$S_{y|\boldsymbol{x}} = AS_{y|z}. \tag{2.2}$$

The central subspace of original data $S_{y|\boldsymbol{x}}$ can be safely reproduced from $S_{y|\boldsymbol{x}} = \Sigma^{-1/2}S_{y|z}$. See Cook and Weisberg (1991), and Cook (1994).

2.1. SIR

As most SDR methods do, SIR requires the linearity condition, $E(z|\eta^T z) = P_\eta z$ where P_η is projection operator for $S_{y|z}$ so that $E(z|\eta^T \boldsymbol{x})$ is linear in z. This condition is satisfied when the distribution of $\boldsymbol{x}$ is elliptically contoured such as multivariate normal distribution. Under (2.1) and the linearity condition

$$E(z|y) = E\left\{E\left(z|\eta^T z\right)|y\right\} = E\left(P_\eta z|y\right).$$

Hence, we have $E(z|y) = p_\eta E(z|y)$. This implies that the inverse regression curve $E(z|y)$ is contained in $S_{y|z}$ since $E(z|\eta^T z)$ is a projection of $E(z|\eta^T z)$ onto $S_{y|z}$. Therefore any $v \in \mathbb{R}^v$, orthogonal to $S_{y|z}$, $E(z|y)$ is degenerate in the direction of v (that is, $v^T\mathrm{Cov}(E(z|y)|v = 0)$. Finding such v is equivalent to finding eigenvectors of $\mathrm{Cov}[E(z|y)]$ whose corresponding eigenvalues are zero. For estimation, a range of y is divided into non-overlapping H slices, $S_1, \ldots, S_H$ when y is continuous variable, $z|y$, so $E(z|y)$ is estimated by averaging z contained in each slice. Slicing is not an approximating procedure if y is discret. The algorithm for SIR is given below.

Step 1. Standardize N observed predictors $\{\boldsymbol{x}_n\}_{i=1}^N$ to obtain $z_i = \hat{\Sigma}^{-1/2}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$, where $\hat{\Sigma}$ and $\bar{\boldsymbol{x}}$ are the sample covariance matrix and the sample mean of $\{\boldsymbol{x}_n\}_{i=1}^N$, respectively.

Step 2. Within each slice $S_1, \ldots, S_H$, compute $\hat{p}_h = \{\sum_{n=1}^n I(y_n \in S_h)\}/N$, and set the mean of $z_i|y_i \in S_h$ by $\bar{z}_h$.

Step 3. Compute the kernel matrix

$$\hat{V}_{z|y} = \sum_{h=1}^H \hat{p}_h(\bar{z}_h - \bar{z})(\bar{z}_h - \bar{z})^T.$$

Step 4. Conduct eigenvalue decomposition on the $\hat{V}_{z|y}$. Let $\hat{\alpha}_1, \ldots, \hat{\alpha}_d$ be the eigenvectors corresponding to d largest eigenvalues.

Step 5. Estimated basis for $S_{y|\boldsymbol{x}}$ becomes $\hat{\beta}_r = \hat{\Sigma}^{-1/2}\hat{\alpha}_r$, $r = 1, \ldots, d$.

The kernel matrix $\hat{V}_{z|y}$ in Step 3 is a sample weighted covariance matrix.

The transformation in Step 1 and Step 5 can be justified by (2.2). The estimator satisfies $\sqrt{n}$ consistency and robust to the number of slices. SIR performs better than other SDR methods for linear structure, but it is usually not true in other cases. It is also known that SIR fails to capture the symmetric structure as discussed in the next section.

2.2. Model-based SIR (Scrucca, 2011)

MSIR was motivated from the inability of SIR to detect a symmetric data structure. Consider 2 predictors x_1 and x_2 with $E(x_i) = 0$ for $i = 1, 2$, and let $y = x_1^2$ without error term for simplicity. The orthonormal basis for the central subspace is $(1, 0)$. But SIR cannot find it because $E(z|y) = 0$ for all y. MSIR is suggested to fix this problem by considering multiple local means of finite GMM. These local means avoid loss of information about symmetric structure as well as allow a flexible approach to handle diverse data structures.

The estimation of MSIR have two main stages as SIR does; computing kernel matrix and conducting eigenvalue decomposition. It also uses slices $S_1, \ldots, S_H$ dividing the range of y. A difference between two methodologies comes from replacing $\bar{z}_h$ with estimated means of components in GMM. MSIR assumes the data in the h^{th} slice have C_h mixture groups that can be described as follows.

$$f(\boldsymbol{x}|y \in S_h) = \sum_{c=1}^{C_h} \lambda_{hc}\phi(\boldsymbol{x} : \mu_{hc},\ \Sigma_{hc}), \tag{2.3}$$

where $\phi(\cdot)$ is the probability density function of multivariate normal distribution with mean μ_{hc} and covariance Σ_{hc}. The λ_{hc} is mixing proportion for the c^{th} mixture group in h^{th} slice, so that $\sum_{c=1}^{C_h} \lambda_{hc} = 1$ and $\lambda_{hc} \leq 0$ for $c = 1, \ldots, C_h$, and $h = 1, \ldots, H$. Under this model, MSIR uses the following weighted kernel matrix,

$$M = \sum_{h=1}^{H}\sum_{c=1}^{C_h} w_{hc}(\mu_{hc} - \bar{\mu})(\mu_{hc} - \bar{\mu})^T, \tag{2.4}$$

where $w_{hc} = p_h\lambda_{hc}$, $p_h = P(y \in S_h)$, and $\bar{\mu} = \sum_h \sum_c p_h\lambda_{hc}\mu_{hc}$.

Selection of covariance structure and the number of components at h^{th} slice C_h can be determined by comparing Bayesian Information Criterion (BIC). The algorithm of MSIR can be described as follows.

Step 1. Divide range of sample response $\{y_n\}_{n=1}^N$ into H slices, $S_1, \ldots, S_H$ and compute $\hat{p}_h = \{\sum_{n=1}^{n} I\ (y_n \in S_h)\}/N$ for $h = 1, \ldots, H$.

Step 2. For each slice, fit GMM as in (2.3). Select the covariance structure $\hat{\Sigma}_{hc}$ and C_h minimizing BIC.

Step 3. Compute the weighted kernel matrix M in (2.4) using $\hat{\mu}_{hc}$ and $\hat{\lambda}_{hc}$ obtained from Step 2.

Step 4. Conduct generalized eigenvalue decomposition of $\hat{M}$ with respect to $\hat{\Sigma} = (1/n)\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\mu})(\boldsymbol{x}_i - \bar{\mu})^T$:

$$\hat{M}\beta_r = \gamma_r\hat{\Sigma}\beta_r\beta_r^T\hat{\Sigma}\beta_s = \begin{cases} 1, & \text{if } r = s, \\ 0, & \text{otherwise} \end{cases}$$

with $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_d > 0$.

The generalized eigenvalue decomposition in Step 4 is identical to finding $B = [\beta_1, \ldots, \beta_d]$ maximizing $B^T MB$, subject to $B^T MB = I$ where I is identity matrix with d diagonal elements. Solutions $\hat{\beta}_1^{MSIR}, \ldots, \hat{\beta}_d^{MSIR}$, resulted from the algorithm above is estimator of MSIR. Scrucca (2011) demonstrated that $\hat{\beta}_1^{MSIR}, \ldots, \hat{\beta}_d^{MSIR}$ span $S_{y|\boldsymbol{x}}$, and they are consistent estimator showing that MSIR can accommodate various data structures.

3. Mixture probabilistic PCA

3.1. Probabilistic PCA (PPCA)

Principal component analysis (PCA) (Jolliffe, 2002) is a technique for dimension reduction in multivariate data sets. Several examples of its many applications include data reduction, pattern recognition, exploratory data analysis and time series prediction. PCA is closely related to factor analysis (Anderson, 1963; Whittle, 1952; Young, 1940), and can be expressed as a latent variable model (Anderson and Rubin, 1956; Lawley, 1953). Further work by Tipping and Bishop (1999a) indicates how PCA may be viewed as a ML procedure based on a probability density model of the observed data. Both of these assume that p dimensional variable has the following structure,

$$\boldsymbol{t} = Wz + \mu + \epsilon,$$

where z is an d-dimensional Gaussian latent variable, W is a $p \times d$ matrix, ϵ and z are independent. However, PPCA uses isotropic noise $\epsilon \sim N(0, \sigma^2 I)$ instead of assuming $\epsilon \sim N(0, \Sigma)$ as in factor analysis. Then, the conditional distribution of $\boldsymbol{t}$ given z is of the form,

$$p(\boldsymbol{t}) = (2\pi)^{-\frac{d}{2}}|C|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{t} - \mu)^T C^{-1}(\boldsymbol{t} - \mu)\right\}, \tag{3.1}$$

where $C = WW^T + \sigma^2 I$. This leads to the log-likelihood of observed data points $\{\boldsymbol{t}_n)_{n=1}^N$,

$$L = -\frac{N}{2}\left\{d\ln(2\pi) - \ln|C| + \mathrm{tr}\left(C^{-1}S\right)\right\},$$

where $S = (1/N)\sum_{i=1}^{N}(\boldsymbol{t}_n - \mu)(\boldsymbol{t}_n - \mu)^T$. Tipping and Bishop (1999a) showed that global maximum of the likelihood occurs when

$$\mu_{ML} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{t}_n,$$
$$W_{ML} = U_d\left(\Lambda_d - \sigma^2 I\right)^{\frac{1}{2}} R, \tag{3.2}$$

where Λ_d is a diagonal matrix whose elements are first d largest eigenvalues of S, and they are arranged in order of decreasing magnitude. U_d is a matrix with i^{th} column as the eigenvector corresponding to the i^{th} largest eigenvalue. R is an arbitrary $d \times d$ orthogonal rotation matrix. Interestingly, at the saddle points of the likelihood function, Λ_d have d largest eigenvalues but they are not orderly comprised; therefore, the same is true of eigenvectors in U_d. Moreover, given $W = W_{ML}$, σ^2 maximizing the likelihood is

$$\sigma^2_{ML} = \frac{1}{p-d} \sum_{j=d+1}^{p} \hat{\lambda}_j,$$

where $\hat{\lambda}_{d+1}, \ldots, \hat{\lambda}_p$ are the $p - d$ smallest eigenvalues of S. As a consequence, columns of W_{ML} span the d-dimensional principal subspace and σ^2_{ML} represents an average variance of the discarded dimensions.

3.2. Mixture PPCA (MPPCA)

The probability model with PCA can be extended to a mixture of local PCA models (Tipping and Bishop, 1999b). The log-likelihood of observed data set for a mixture model is

$$L = \sum_{n=1}^{N} \ln[p(\boldsymbol{t}_n)] = \sum_{n=1}^{N} \ln \left\{ \sum_{i=1}^{M} \pi_i p(\boldsymbol{t}_n|i) \right\}.$$

The $p(\boldsymbol{t}_n|i)$ is the probability density function of the i^{th} mixture components single PPCA model and π_i is a mixing proportion for the i^{th} component with $\pi_i \geq 0$ for $i = 1, \ldots, M$ and $\sum_{i=1}^{M} \pi_i = 1$. The form of the likelihood is identical to the standard finite Gaussian mixture model with covariance structures,

$$C_i = W_i W_i^T + \sigma_i^2 I. \tag{3.3}$$

R_{ni}, posterior probability or responsibility for i^{th} observation, can be expressed as

$$R_{ni} = \frac{p(\boldsymbol{t}_n|i)\pi_i}{p(\boldsymbol{t}_n)}.$$

In the EM algorithm, both μ_i and π_i are updated as those in the standard GMM. After that, W_i and σ_i^2 are determined by the standard eigenvalue decomposition of weighted covariance matrix,

$$S_i = \frac{1}{\tilde{\pi}_i N} \sum_{n=1}^{N} R_{ni}(\boldsymbol{t}_n - \tilde{\mu}_i)(\boldsymbol{t}_n - \tilde{\mu}_i)^T$$

in the same manner of a single PPCA. Eventually, each W_i spans the local principal subspace at the convergence of the algorithm.

MPPCA clusters observations in a soft manner. To investigate its clustering mechanism, recall that the posterior probability R_{ni} is proportional to probability density function of the i^{th} component

$$R_{ni} \propto p(\boldsymbol{t}_n|i) \propto |C_i|^{-\frac{1}{2}} \exp\left\{ -\frac{E_{ni}^2}{2} \right\},$$

where $E_{ni}^2 = (\boldsymbol{t}_n - \mu_i)^T C_i^{-1} (\boldsymbol{t}_n - \mu_i)$. At the maximum of likelihood, from the Equations (3.2) and (3.3), E_{ni}^2 can be derived as

$$E_{ni}^2 = (\boldsymbol{t}_n - \mu_i)^T \left\{ \sigma_i^2 I + U_{qi} \left(\Lambda_{qi} - \sigma_i^2 I \right) U_{qi}^T \right\}^{-1} (\boldsymbol{t}_n - \mu_i),$$

and defining $z_n = (\boldsymbol{t}_n - \mu_i)$ and $D_i = \text{diag}(\lambda_1, \ldots, \lambda_{qi}, \sigma_i^2, \ldots, \sigma_i^2)$ gives

$$E_{ni}^2 = z_n^T D_i^{-1} z_n.$$

E_{ni}^2 can therefore be interpreted as a weighted sum of squared errors by projecting inside and outside of the i^{th} components principal subspace spanned by W_i. Therefore, MPPCA clusters data points by proximity to each subspaces as well as by reconstruction error $\sum_{i=1}^N ||\boldsymbol{t}_n - \hat{\boldsymbol{t}}_n||^2$, where $\hat{\boldsymbol{t}}_n$ is projected n^{th} observation to the estimated subspace.

Tipping and Bishop (1999b, Appendix C) provided iterative EM algorithm for MPPCA, and also proposed its simplified version, which guarantees to find the local maximum of expected complete-data log likelihood.

4. Model-based SIR for mixtures of regression models

In the regression problem with response variable y and p dimensional predictors $\boldsymbol{x}$, suppose that y was drawn from a population which consists of M groups or components with proportions $\pi_1, \ldots, \pi_M$. Let G be a categorical variable whose value ranges from 1 to M, and indicates the group y belongs to, so that $\Pr(G = i) = \pi_i$ and $\sum_{i=1}^M \pi_i = 1$. If each of the i^{th} group has d_i dimensional central subspace, then $(y, \boldsymbol{x})$ has the following property,

$$y \perp\!\!\!\perp \boldsymbol{x} | B_i^T \boldsymbol{x}, G = i, \quad \text{for } i = 1, \ldots, M, \tag{4.1}$$

where B_i is d_i by p matrix whose columns span the central subspace for the i^{th} group denoted as $S_{y|\boldsymbol{x},G=i}$. We further assume that $S_{y|\boldsymbol{x},G=i} \not\subset S_{y|\boldsymbol{x},G=i}$ if $i \neq j$ in order to consider only cases with distinct subspaces.

Our method is mainly composed of two parts. We estimate within-slice local means by finite GMM first, and then we exploit MPPCA to estimate M local principal subspaces of each mixture group. At the first stage, the distribution of $\boldsymbol{x}|y$ is approximated by fitting finite GMM in (2.2) for slices of y. Scrucca (2011) showed that d eigenvectors of M in (2.4) span the central subspace. Observations within a slice came from more than one group and if data structures of all those groups are distinguishable by several Gaussian distributions, $\hat{\mu}_{hc}$ would capture the locational information about $(\boldsymbol{x}|y, G = i)$ for $i = 1, \ldots, M$. Putting all $\hat{\mu}_{hc}$ together in p dimensional space, these would lie on M distinct subspaces $S_{y|\boldsymbol{x},G=i}$.

In the second part, we estimate the M central subspaces by MPPCA regarding $\hat{\mu}_{hc}$ as multivariate observations for MPPCA. This gives complete-data log likelihood

$$L = \sum_{n=1}^{n'} \ln p(\hat{\mu}_{hc}) = \sum_{n=1}^{n'} \ln \{\pi_i p(\hat{\mu}_{hc}|i)\},$$

where $p(\cdot|i)$ is the probability density function of the i^{th} component's single PPCA model as described in Section 3. A simplified version of Iterative EM algorithm described in Tipping and Bishop (1999b, Appendix C) can be used for MPPCA. At each iteration, we give weights for $\hat{\mu}_{hc}$ by multiplying $\hat{w}_{hc} = \hat{p}_h \hat{\lambda}_{hc}$ to posterior probability as if $\hat{\mu}_{hc}$ was observed as many time as $\hat{w}_{hc}$. Finally, under the assumption that the number of mixture components M and dimension of principal subspace d_i $(i = 1, \ldots, M)$ were presumed to be known, we may form the following algorithm.

Step 1. Standardize N observed predictors $\{\boldsymbol{x}\}_{n=1}^N$ to obtain $z_i = \hat{\Sigma}^{-1/2}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$, where $\hat{\Sigma}$ and $\bar{\boldsymbol{x}}$ are the sample covariance matrix and the sample mean of $\{\boldsymbol{x}_n\}_{n=1}^N$ respectively.

Step 2. Divide range of sample response $\{y_n\}_{i=1}^N$ into H slices, $S_1, \ldots, S_H$ and compute $\hat{p}_h = \{\sum_{i=1}^N I$ $(y_n \in S_h)\}/N$ for $h = 1, \ldots, H$.

Step 3. For the h^{th} slice, fit GMM, and let $\hat{\mu}_{hc}$ be the estimated mean of c^{th} mixture component. Collect $\hat{\lambda}_{hc}$ corresponding to $\hat{\mu}_{hc}$ for $c = 1, \ldots, C_h$ and $h = 1, \ldots, H$. The covariance structures of Σ_{hc} and the number of mixture components in the h^{th} slice C_h minimizing BIC are selected.

Step 4. Update parameters for MPPCA in a sequence of R_{hci}, π_i, μ_i, S_i, W_i, σ_i^2 as:

$$\begin{aligned}
R_{hci} &= \frac{p(\hat{\mu}_{hc}|i)\pi_i}{p(\hat{\mu}_{hc})},\\
\pi_i &= \sum_{h=1}^{H}\sum_{c=1}^{C_h} \hat{w}_{hc} R_{hci},\\
\bar{\mu}_i &= \frac{\sum_{h=1}^{H}\sum_{c=1}^{C_h} R_{hci}\hat{w}_{hc}\hat{\mu}_{hc}}{\sum_{h=1}^{H}\sum_{c=1}^{C_h} R_{hci}\hat{w}_{hc}},\\
S_i &= \frac{\sum_{h=1}^{H}\sum_{c=1}^{C_h} R_{hci}\hat{w}_{hc}\hat{\mu}_{hc}(\hat{\mu}_{hc}-\bar{\mu}_i)(\hat{\mu}_{hc}-\bar{\mu}_i)^T}{\sum_{h=1}^{H}\sum_{c=1}^{C_h} R_{hci}\hat{w}_{hc}},\\
\bar{W}_i &= S_i W_i \left(\sigma_i^2 I + M_i^{-1} W_i^T S_i W_i\right)^{-1},\\
\bar{\sigma}_i^2 &= \frac{1}{d}\operatorname{tr}\left(S_i - S_i W_i M_i^{-1} \bar{W}_i^T\right).
\end{aligned}$$

where $\hat{w}_{hc} = \hat{p}_h\hat{\lambda}_{hc}$, and μ_i is the i^{th} mixture component's mean in MPPCA. Note that $\bar{W}_i$ and $\bar{\mu}_i$ are newly updated one while W_i and $\hat{\mu}_i$ are previous estimators.

Step 5. Iterate Step 4 until it converges.

Step 6. Estimator $\hat{B}_i$ for the basis of $S_{y|\mathbf{x},G=i}$ can be obtained by transforming $\hat{W}_i$ into $\hat{\Sigma}^{-1/2}\hat{W}_i$.

4.1. Inherent problems in MPPCA

Since we are using MPPCA as a tool of subspace clustering, we have the same kind of problems innate in MPPCA. One of the problems is that M and d_i $(i = 1, \ldots, M)$ should be previously known or given. According to Vidal (2011), by giving priors to parameters in MPPCA the problem can be addressed as shown in Paisley and Carin (2009). All algorithms in this paper are based on the assumption that M and d_is are given.

Another problem of MPPCA is that it frequently converges to local maximum as usual in GMM so that estimates are highly dependent on initial values. Starting with multiple initial values is the general approach, but the parameter space of MPPCA is considerably extensive especially because of W_is. This may require significant computing time and is not a practical option. Spurious solutions also add difficulties in finding a global maximum. Seo and Kim (2012) suggested k-deleted likelihood criteria. Computing likelihood without k influential or divergent likelihood terms, a k-deleted log likelihood provides simple criteria for root selection, avoiding spurious roots. The MLE selected by this criteria satisfies the consistency.

We suggest using standard PCA results as initial values of W_i, and will show that they are satisfactory to estimate true parameters in (4.1) by simulation studies. When N is large, one of our initial values of W_i tend to have the largest k-deleted log likelihood value with a rapid convergence to true parameters that are also robust to the number of slices.

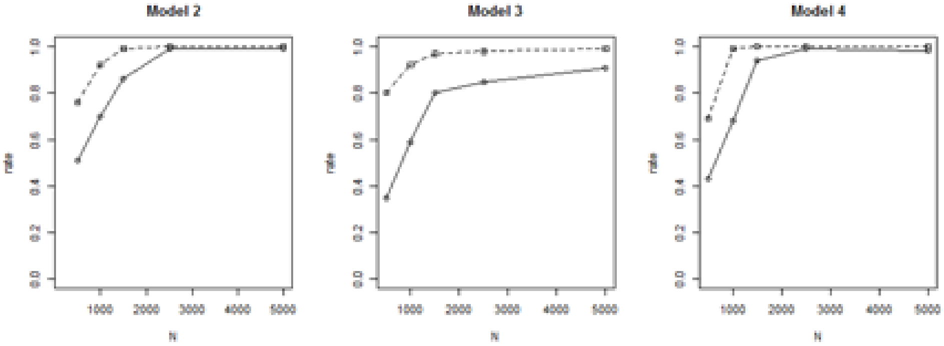

Figure 1: *The rates that W^{Init*} converges to the highest local maximum (solid line), and the rates that W^{Init*} converges to one of two highest local maximum (dotted line).*

4.2. Initial values and root selection

Under the assumption that the number of components M and total number of dimensions $D = \sum_i d_i$ ($i = 1, \ldots, M$) are given, we can start with initial values of W_i by considering D principal components obtained by standard PCA on $\hat{\mu}_{hc}$, and let $v_1, \ldots, v_D$ be those D principal components. Under (4.1), $E(z|G) = 0$ for all G, and $\text{span}\{v_1, \ldots, v_D\} \supset \cup_{i=1}^{M} S_{y|z,G=i}$. Hence, all structural information about $(y, \boldsymbol{x})$ would be contained in $\text{span}\{v_1, \ldots, v_D\}$, and any basis of each $S_{y|\boldsymbol{x}.G=i}$ would be a linear combination of $v_1, \ldots, v_D$. Once we assort D dimensional space $\text{span}\{v_1, \ldots, v_D\}$ into M subspaces whose dimension are $d_1, \ldots, d_M$, MPPCA algorithm would automatically rotate and fit the subspaces to the data points. Thus, it is adequate to consider all possible combinations of grouping D dimensional space into M subspaces as initial values for W_i.

However, still the root selection is another problem. We conducted a simulation study to check convergence of our initialization method. In order to set initial values of W_i, we grouped $v_1, \ldots, v_D$ into M sets whose sizes are $d_1, \ldots, d_M$, as described above. Let $W^{Init} = (W_1^{Init}, \ldots, W_M^{Init})$ be one case of this grouping. We fixed initial values of π_i, σ_i, and μ_i ($i = 1, \ldots, M$) for all simulated data sets. Let W^{Init*} be the closest W^{Init} to true parameters such that

$$W^{Init*} = \arg\min_{W^{Init}} \left\{ \sum_{i=1}^{M} \left\| P_{S_{y|\boldsymbol{x},G=i}} - P_{C\left(W_i^{Init}\right)} \right\|_2 \right\},$$

where $C(W_i^{Init})$ is the column space of W_i^{Init}. Solid lines in Figure 1 are rates that W^{Init} converges to the highest local maximum among all possible combinations of initial values during 100 replications. The dotted lines in Figure 1 are rates that W^{Init*} converges to one of two highest local maximum. Since Model 1 has just one case of W^{Init}, results of Model 2, 3, and 4 are included only (Section 5.1 provides details of the models). To avoid spurious solutions, we computed heights of local maximum by using k-deleted log likelihood criteria with $k = 6$. As N increases, W^{Init*} converges to the highest local maximum in all 3 models.

5. Simulation study

We use simulations to examine the performance of MPPCA in recovering true subspaces as well as compare performance in terms of several measures for estimation accuracy and classification ability. The following sections will introduce the details of the considered models and measures.

5.1. Models

Four mixture models are considered for the simulation studies. All models consist of 2 mixture groups with different structures. Model 1 contains two components whose central subspaces are one dimensional. The former group has a symmetric structure that SIR cannot detect; however, the latter group has a nearly monotone and nonlinear structure. Both of two groups in Model 2 have a nonlinear structure but a different central subsapce dimension. In Model 3, variance of the response changes along with $\boldsymbol{x}^T\beta_3$ axis; in addition, both the first and the second group have a different dimensional data structure. Model 4 consist of 2 components that have two dimensional central subspaces. We pick unequal probabilities 0.7 and 0.3 for each group with no special purpose. We tried several additional equal and unequal probability sets in the simulation. However, there were no significant differences and we only report results for 0.7 and 0.3 cases.

- Model 1:

$$y = \begin{cases} \left(\beta_1^T \boldsymbol{x}\right)^2 + \sigma\epsilon, & \text{with probability 0.7,} \\ \sin\left(\beta_2^T \boldsymbol{x}\right) + \sigma\epsilon, & \text{with probability 0.3.} \end{cases}$$

- Model 2:

$$y = \begin{cases} \dfrac{\beta_1^T \boldsymbol{x}}{0.5 + \left(\beta_2^T \boldsymbol{x} + 1.5\right)^2} + \sigma\epsilon, & \text{with probability 0.7,} \\ -5 + \left(\beta_3^T \boldsymbol{x}\right)^2 + \sigma\epsilon, & \text{with probability 0.3.} \end{cases}$$

- Model 3:

$$y = \begin{cases} \dfrac{1}{2}\left[1 + \left(\beta_3^T \boldsymbol{x}\right)^2\right]\epsilon, & \text{with probability 0.7,} \\ -1 + 0.4\left(\beta_1^T \boldsymbol{x}\right)^2 + \left|\beta_2^T \boldsymbol{x}\right|^{\frac{1}{2}} + \sigma\epsilon, & \text{with probability 0.3.} \end{cases}$$

- Model 4:

$$y = \begin{cases} \dfrac{\left(\beta_1^T \boldsymbol{x}\right)}{0.5 + \left(\beta_2^T \boldsymbol{x} + 1.5\right)^2} + \sigma\epsilon, & \text{with probability 0.7,} \\ -3 + \left(\beta_3^T \boldsymbol{x}\right)\left(\beta_4^T \boldsymbol{x}\right) + \sigma\epsilon, & \text{with probability 0.3.} \end{cases}$$

Six dimensional β_i has 1 at the i^{th} position and 0s otherwise. Predictors and ϵ come from independent standard normal distribution.

5.2. Accuracy of estimation

The notion of a maxim angle was used to measure the angle between estimated and true central subspaces. For any non-zero subspaces A and B , let P_A and P_B are the orthogonal projectors onto A and B respectively. Then θ $(0 \leq \theta \leq \pi/2)$ is defined as maximal angle such that $||P_A - P_B||_2 = \sin\theta$ where $|| \cdot ||$ is the spectral Euclidean norm (Gentle, 2007; Meyer, 2000, p.455).

Table 1: Means and standard deviations of maximal angles between true and estimated central subspaces for Model 1

	$N = 500$		$N = 1,500$		$N = 2,500$		$N = 5,000$	
	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2
5 slices	38.515	10.562	12.283	4.555	6.700	3.387	4.580	2.341
(sd)	16.119	3.817	7.945	1.484	2.398	0.993	1.958	0.757
7 slices	27.211	9.432	12.630	4.866	6.913	3.280	3.925	2.223
(sd)	15.109	3.453	7.229	1.541	2.757	0.970	1.380	0.711
9 slices	21.013	9.394	13.907	4.646	7.348	3.286	4.569	2.135
(sd)	9.700	2.866	10.026	1.703	3.012	1.123	1.776	0.733
13 slices	29.069	12.156	10.273	4.490	6.914	3.383	4.059	2.168
(sd)	19.883	4.278	4.959	1.529	2.556	1.018	1.328	0.649
15 slices	30.322	13.829	10.268	4.980	7.667	3.485	4.143	2.335
(sd)	16.836	9.350	5.451	1.603	3.237	1.375	1.549	0.691

Table 2: Means and standard deviations of maximal angles between true and estimated central subspaces for Model 2

	$N = 500$		$N = 1,500$		$N = 2,500$		$N = 5,000$	
	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2
5 slices	73.899	17.071	36.050	8.460	10.606	5.869	4.990	3.568
(sd)	19.713	6.823	29.792	2.923	11.221	2.110	1.858	0.973
7 slices	64.125	16.148	9.766	6.947	6.296	5.038	4.140	3.444
(sd)	24.578	6.571	3.801	2.153	1.894	1.313	1.366	0.855
9 slices	47.235	16.326	9.322	7.085	5.993	5.277	3.846	3.222
(sd)	23.871	7.160	3.823	2.327	1.845	1.474	1.310	0.922
13 slices	45.746	21.133	9.491	6.848	5.620	4.975	3.655	3.461
(sd)	23.011	12.694	8.527	2.023	1.916	1.300	1.187	1.046
15 slices	35.107	23.474	10.501	7.499	5.552	5.542	3.857	3.388
(sd)	23.927	13.637	5.908	2.188	1.856	1.333	1.305	0.897
18 slices	34.445	24.245	10.969	7.528	6.071	5.593	3.807	3.389
(sd)	22.925	14.503	8.751	2.724	3.372	1.723	1.240	0.933
20 slices	36.580	23.626	12.599	7.675	6.663	5.404	4.054	3.341
(sd)	23.489	14.833	11.993	2.118	3.523	1.586	1.300	0.970

Tables 1–4 report mean and standard deviation of maximal angles between true and estimated central subspaces for Models 1–4 after 100 replications with $\sigma = 0.1$, respectively. It shows that our method fails to detect true subspaces in all cases when the number of slices and the number of observations are small. In addition, the angles become worse when the number of dimensions in each group increases. We can see cases with larger than 70 degrees in maximal angle with 5 and 7 slices for Model 2, 3, and 4. However, the maximal angle decreases to one digit in almost all cases when the number of observations is larger than 2,500 and number of slices is larger than 8.

Tables 5 and 6 include estimated mixing proportions for the second groups in the 4 models. The estimated mixing proportions for the second group converges to the true value 0.3 except for cases with $N = 500$ in Model 1. Small number of slices provide relatively large standard deviation estimates when estimating mixing proportions.

5.3. Classification

Our MPPCA algorithm contains two kinds of posterior probabilities, one from the GMM of MSIR, and the other from a mixture part of MPPCA. For each $h = 1, \ldots, H$, $\hat{\mu}_{h1}, \ldots, \hat{\mu}_{hC_h}$ represent local means within h^{th} slice with $\boldsymbol{x}_i|y_i \in S_h$. Let us denote this first posterior probability as $P(c = c'|[\boldsymbol{x}_i|y_i] \in$

Table 3: Means and standard deviations of maximal angles between true and estimated central subspaces for Model 3

	N = 500		N = 1,500		N = 2,500		N = 5,000	
	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2
5 slices	83.844	60.963	13.367	7.816	9.010	5.537	10.285	3.607
(sd)	15.237	34.775	9.745	4.616	2.700	1.862	18.449	0.915
7 slices	70.567	21.004	12.438	7.533	8.118	5.173	5.282	3.290
(sd)	27.695	18.849	5.860	2.605	2.648	1.398	1.320	0.913
9 slices	50.364	15.623	10.574	6.208	7.942	4.891	5.330	3.356
(sd)	32.662	6.358	3.231	1.448	2.160	1.336	1.672	0.950
13 slices	37.372	19.618	10.286	6.982	7.613	4.934	4.841	3.288
(sd)	17.314	9.344	3.084	2.061	2.043	1.414	1.219	0.983
15 slices	36.490	22.259	10.207	6.752	7.274	4.795	4.587	3.219
(sd)	15.601	10.111	2.731	2.063	2.038	1.372	1.188	0.909
18 slices	38.157	21.009	10.168	6.741	7.578	5.246	4.650	3.247
(sd)	16.706	8.435	3.298	1.933	2.119	1.508	1.129	0.916
20 slices	39.501	22.364	10.541	7.171	7.360	4.890	4.726	3.166
(sd)	14.035	8.443	3.471	2.471	1.885	1.196	1.247	0.881

Table 4: Means and standard deviations of maximal angles between true and estimated central subspaces for Model 4

	N = 500		N = 1,500		N = 2,500		N = 5,000	
	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2
5 slices	80.460	57.006	69.624	6.023	61.916	4.384	23.772	2.959
(sd)	15.315	19.953	17.753	4.716	23.043	1.446	28.757	1.028
7 slices	71.797	53.663	58.974	5.308	34.518	4.013	5.235	2.814
(sd)	16.959	23.083	23.457	1.838	31.202	1.300	1.488	0.912
9 slices	66.047	57.561	31.100	5.366	11.281	3.825	4.998	2.639
(sd)	20.707	19.637	26.851	1.674	11.414	1.176	1.428	0.830
13 slices	54.467	52.984	18.735	5.420	7.674	3.982	5.643	2.720
(sd)	22.462	22.390	17.878	1.853	2.161	1.169	1.631	0.811
15 slices	61.679	59.029	19.123	5.380	7.659	3.963	5.629	2.553
(sd)	20.527	21.591	17.285	1.691	2.149	1.275	1.611	0.957
18 slices	58.246	58.175	18.920	5.421	8.099	4.014	5.269	2.725
(sd)	22.240	23.247	16.015	1.859	2.200	1.370	1.563	0.793
20 slices	53.481	55.832	20.541	5.497	8.192	3.933	5.050	2.721
(sd)	21.141	22.336	16.085	1.712	2.875	1.332	1.406	0.911

Table 5: Estimated mixing proportions of the second group in Model 1 and 2

	Model 1				Model 2			
	N = 500	N = 1,500	N = 2,500	N = 5,000	N = 500	N = 1,500	N = 2,500	N = 5,000
5 slices	0.611	0.330	0.298	0.322	0.310	0.289	0.248	0.238
(sd)	0.133	0.140	0.105	0.107	0.113	0.114	0.100	0.076
7 slices	0.609	0.384	0.349	0.330	0.221	0.226	0.271	0.285
(sd)	0.120	0.090	0.080	0.063	0.097	0.078	0.055	0.048
9 slices	0.672	0.385	0.326	0.324	0.205	0.223	0.278	0.293
(sd)	0.128	0.092	0.067	0.054	0.114	0.063	0.055	0.041
13 slices	0.574	0.452	0.357	0.299	0.270	0.239	0.273	0.300
(sd)	0.152	0.091	0.063	0.046	0.116	0.043	0.034	0.029
15 slices	0.597	0.504	0.367	0.312	0.330	0.237	0.268	0.306
(sd)	0.134	0.080	0.066	0.055	0.131	0.052	0.035	0.028
18 slices	0.555	0.523	0.397	0.318	0.359	0.237	0.260	0.302
(sd)	0.135	0.082	0.063	0.052	0.124	0.061	0.042	0.027
20 slices	0.551	0.560	0.419	0.312	0.385	0.231	0.250	0.292
(sd)	0.131	0.078	0.065	0.041	0.127	0.063	0.034	0.033

Table 6: Estimated mixing proportions of the second group in Model 3 and 4

	Model 1				Model 2			
	$N = 500$	$N = 1,500$	$N = 2,500$	$N = 5,000$	$N = 500$	$N = 1,500$	$N = 2,500$	$N = 5,000$
5 slices	0.550	0.324	0.329	0.314	0.497	0.549	0.476	0.345
(sd)	0.207	0.044	0.030	0.037	0.147	0.176	0.176	0.168
7 slices	0.284	0.327	0.337	0.331	0.448	0.428	0.327	0.285
(sd)	0.157	0.041	0.018	0.012	0.182	0.175	0.175	0.085
9 slices	0.210	0.226	0.231	0.231	0.300	0.284	0.268	0.270
(sd)	0.091	0.018	0.014	0.012	0.160	0.132	0.045	0.054
13 slices	0.218	0.202	0.243	0.298	0.250	0.261	0.253	0.279
(sd)	0.088	0.034	0.034	0.020	0.124	0.129	0.033	0.054
15 slices	0.236	0.229	0.249	0.266	0.247	0.238	0.261	0.291
(sd)	0.093	0.034	0.020	0.014	0.105	0.113	0.041	0.035
18 slices	0.230	0.220	0.231	0.262	0.255	0.222	0.253	0.285
(sd)	0.086	0.024	0.020	0.014	0.104	0.122	0.042	0.034
20 slices	0.242	0.210	0.226	0.246	0.314	0.189	0.250	0.292
(sd)	0.086	0.023	0.016	0.019	0.112	0.112	0.041	0.028

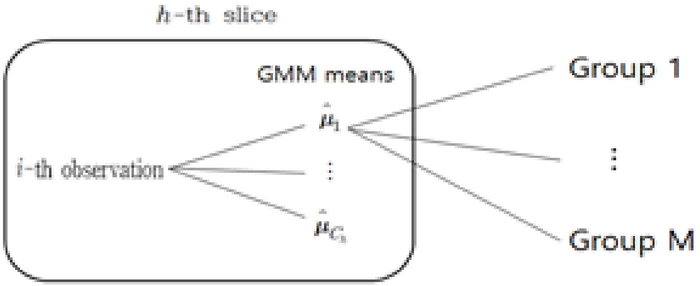

Figure 2: *The mechanism of computing posterior probability as described in Section 5.3*

S_h). Other posterior probabilities come from the MPPCA that clusters $\hat{\mu}_{hc}$ into M Gaussian distributions whose principal subspaces are $S_{y|\boldsymbol{x},G=g}$ $(g = 1, \ldots, M)$. Likewise, each $\hat{\mu}_{hc}$ are responsible for $S_{y|\boldsymbol{x},G=g}$ with probability $P(G = g|\hat{\mu}_{hc})$. Thus, we can calculate combined posterior probability that the i^{th} observation belongs to the group G by

$$\sum_{c=1}^{C_h} P(G = g|\hat{\mu}_{hc'}) \times P(c = c'|[\boldsymbol{x}_i|y_i] \in S_h), \quad \text{for } g = 1, \ldots, M.$$

Figure 2 shows the mechanism of calculating posterior probabilities.

Figure 3 is a scatter plot of 5,000 simulated data points from the following mixture model with 3 groups,

$$y = \begin{cases} -2 + \left(\beta_1^T \boldsymbol{x} - 1\right)^2 + \sigma\epsilon, & \text{with } \Pr(G = 1), \\ 2 + \dfrac{0.5\beta_3^T \boldsymbol{x}}{0.5 + \left(\beta_4^T \boldsymbol{x} + 1.5\right)^2} + \sigma\epsilon, & \text{with } \Pr(G = 2), \\ 0.5 \sin\left(\beta_2^T \boldsymbol{x}\right) + \sigma\epsilon, & \text{with } \Pr(G = 3), \end{cases}$$

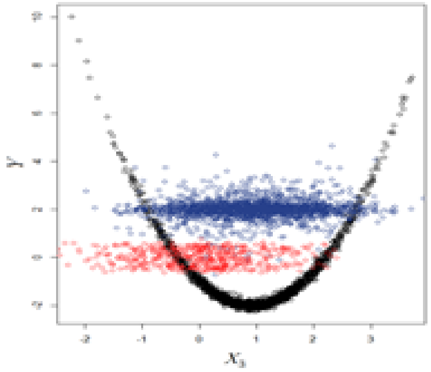

Figure 3: *Scatter plot of 5,000 points from the first (black), second (blue), and third (red) component of the mixture model (5.1).*

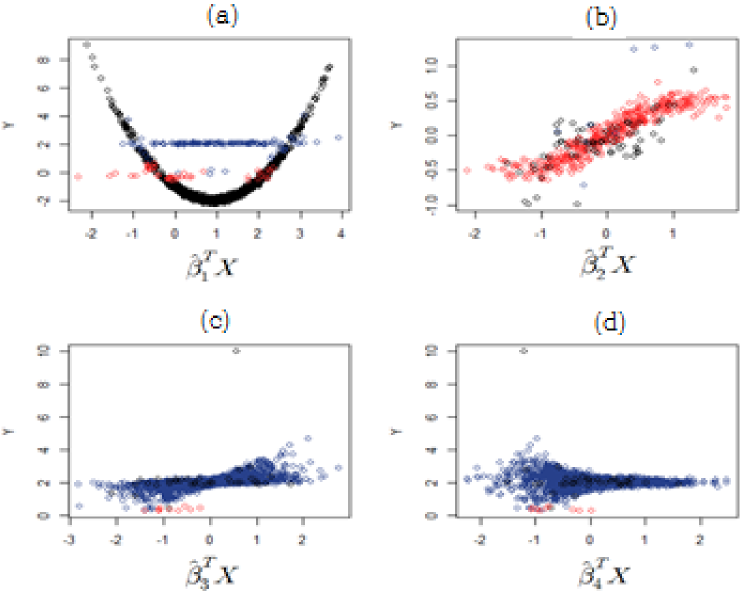

Figure 4: *Plot (a) and (b) contain the points that classified to the first and third group. Those in (c) and (d) are classified to the second group.*

where $\Pr(G = i) = \binom{2}{i-1}(1/3)^{i-1}(2/3)^{3-i}$ for $i = 1, 2, 3$, and $\sigma = 0.1$. Points are colored according to their true mixture group. From the first to the third groups, their samples were colored with black, blue, and red respectively. To clearly distinguish those 3 groups through the plot, we sampled $\boldsymbol{x}$ from different distributions. For the first and second groups, $\boldsymbol{x}$ is generated from $MVN(m, I)$ with $m = (0, 0, 1, 0, 0, 0)$, and $MVN(0, I)$ for the third group with all errors from standard normal distribution.

We computed posterior probabilities of individual points, and classified them into one of three groups with the highest posterior probability. In Figure 4, points classified to the first group are displayed in (a), and points classified to the third group are plotted in (b). Plot of Figure 4(c) and (d) are the scatter plot of the second group whose horizontal axis are the estimated basis of two central

Table 7: Mean and standard deviation of classification error rates for Model 1 and 2

	Model 1				Model 2			
	$N = 500$	$N = 1,500$	$N = 2,500$	$N = 5,000$	$N = 500$	$N = 1,500$	$N = 2,500$	$N = 5,000$
5 slices	0.500	0.258	0.190	0.196	0.225	0.182	0.188	0.101
(sd)	0.106	0.101	0.083	0.091	0.121	0.090	0.108	0.092
7 slices	0.408	0.212	0.149	0.118	0.210	0.167	0.110	0.078
(sd)	0.109	0.080	0.075	0.062	0.112	0.078	0.070	0.060
9 slices	0.426	0.237	0.164	0.115	0.252	0.157	0.121	0.070
(sd)	0.097	0.068	0.065	0.059	0.111	0.071	0.063	0.042
13 slices	0.457	0.272	0.193	0.132	0.359	0.130	0.102	0.065
(sd)	0.088	0.057	0.057	0.051	0.112	0.048	0.049	0.030
15 slices	0.464	0.302	0.211	0.124	0.399	0.147	0.097	0.073
(sd)	0.080	0.046	0.059	0.048	0.087	0.044	0.043	0.028
18 slices	0.458	0.323	0.237	0.146	0.406	0.164	0.108	0.072
(sd)	0.093	0.047	0.057	0.050	0.091	0.050	0.045	0.028
20 slices	0.455	0.340	0.259	0.132	0.427	0.186	0.123	0.077
(sd)	0.082	0.053	0.049	0.040	0.097	0.054	0.045	0.028

Table 8: Mean and standard deviation of classification error rates for Model 3 and 4

	Model 1				Model 2			
	$N = 500$	$N = 1,500$	$N = 2,500$	$N = 5,000$	$N = 500$	$N = 1,500$	$N = 2,500$	$N = 5,000$
5 slices	0.349	0.052	0.040	0.047	0.438	0.438	0.335	0.230
(sd)	0.145	0.050	0.017	0.043	0.200	0.200	0.147	0.130
7 slices	0.197	0.052	0.028	0.022	0.471	0.471	0.254	0.112
(sd)	0.101	0.033	0.008	0.007	0.199	0.199	0.128	0.063
9 slices	0.190	0.133	0.129	0.125	0.355	0.355	0.144	0.139
(sd)	0.074	0.007	0.008	0.011	0.180	0.180	0.061	0.038
13 slices	0.256	0.148	0.100	0.030	0.306	0.306	0.126	0.112
(sd)	0.058	0.034	0.039	0.021	0.113	0.113	0.032	0.046
15 slices	0.287	0.113	0.092	0.081	0.322	0.322	0.125	0.100
(sd)	0.066	0.031	0.013	0.004	0.100	0.100	0.038	0.032
18 slices	0.288	0.127	0.113	0.079	0.301	0.301	0.134	0.095
(sd)	0.055	0.028	0.021	0.015	0.095	0.095	0.033	0.026
20 slices	0.294	0.138	0.117	0.098	0.329	0.329	0.135	0.100
(sd)	0.055	0.023	0.012	0.019	0.112	0.112	0.041	0.028

subspaces. The maximal angles between true and estimated central subspaces are less than 5 degrees in all three groups and estimated mixing proportions are also close to the true values. Except that some points from group two are incorrectly assigned to the first group, most of points have few problems to find their home group.

Table 7 and 8 include the mean and standard deviation of the 100 estimated classification error rates for Model 1, 2, 3, and 4. Classification error rates decrease as N increases; however, the number of slices has little effect on the overall error rates. Most classification errors come from the second part of the MPPCA algorithm, i.e. mixture part for PPCA modeling.

6. Real data example

We consider Australian Institute of Sports (AIS) data set on 102 male and 100 female athletes collected by the AIS, courtesy of Richard Telford and Ross Cunningham (Cook and Weisberg, 1994). The variable lean body mass (LBM) is set to be response and includes only two variables, sum of skin folds (SSF), and weight in kg (Wt) as predictors. We intentionally delete the variable sex in the model

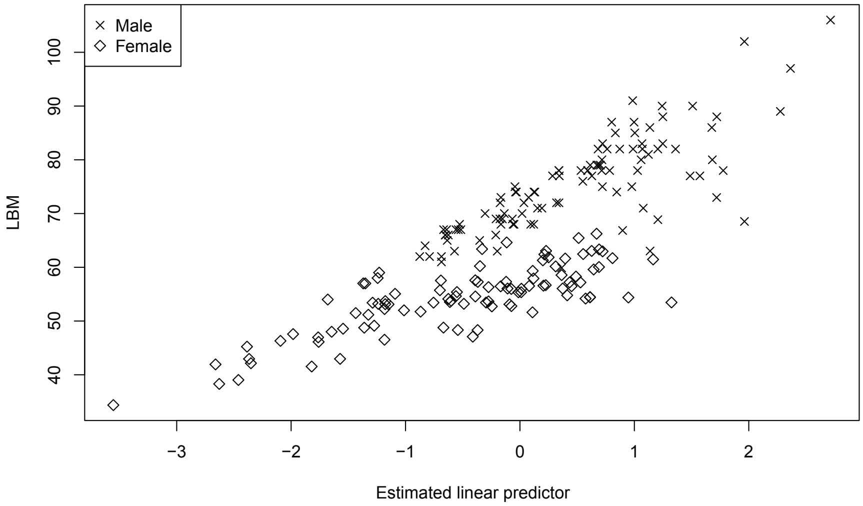

Figure 5: *Plot of estimated first linear predictors vs. lean body mass (LBM).*

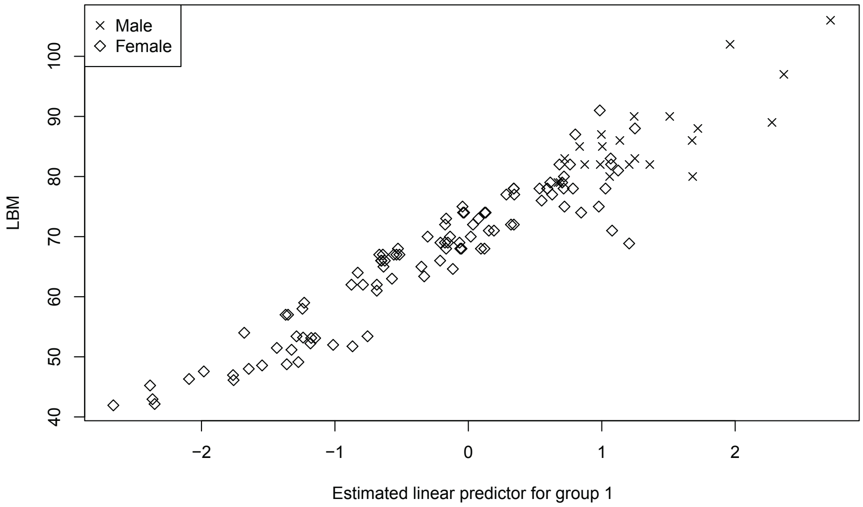

Figure 6: *Plot of estimated first linear predictors vs. lean body mass (LBM) in group 1.*

and check classification errors and other aspects of our method compared with other inverse regression methods assuming that the number of group is one.

Permutation test and BIC type criterion adapted for MSIR (Scrucca, 2011) that selected only one dimension is adequate with a 5% significance level. However, Figure 5 shows two distinct patterns, one for male athletes and the other for females. MPPCA assuming $M = 2$ and one dimensional component for each group divide males and females into two separate groups with a classification error rate 0.19 and reveals strong linear pattern for each group as indicated in Figures 6 and 7.

7. Concluding remarks

This paper proposes a SDR method for mixture data. We assume that the data contains M mixture

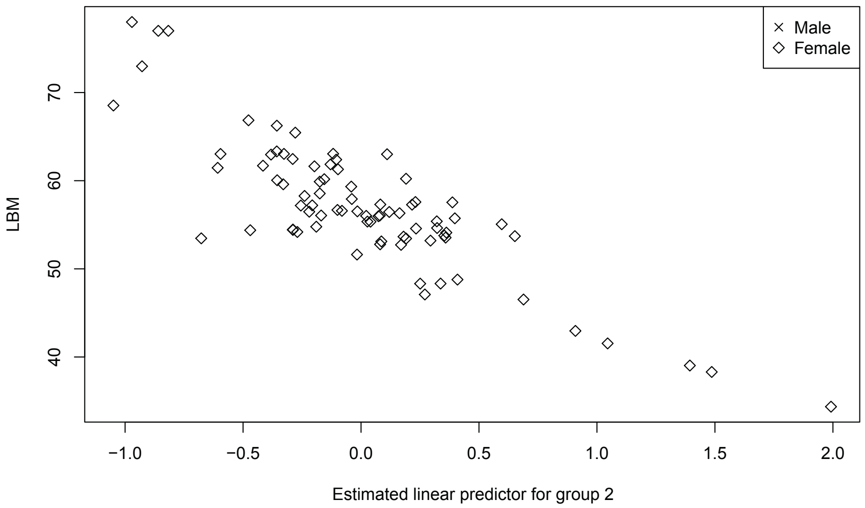

Figure 7: *Plot of estimated first linear predictors vs. lean body mass (LBM) in group 2.*

components with different central subspaces. We combined the MSIR and MPPCA approach in a simple and intuitive way to estimate the basis of central subspaces and mixing proportions for fixed M and a number of components. We feed estimated means for each slice from finite GMM into MPPCA to cluster them into M subspaces. Simulations indicate that our algorithm using k-deleted log-likelihood criterion is successful to search appropriate subspaces and estimate converges to true parameter values.

It is possible to extend likelihood ratio tests and EM tests for GMM (Chen *et al.*, 2012; Jeffries, 2003; Lo, 2005; Lo *et al.*, 2001) and apply it sequentially when selecting the number of groups and total number of components. In addition, we may also try several possible combinations of groups and components and select the best model by cross validation or comparing log-likelihood values.

The method proposed in this paper can be applied to regression problems in which data set comes from several groups and usual linear regression modeling with only one linear predictor that fails in capturing complex regression structures inside.

References

Anderson TW (1963). Asymptotic theory for principal component analysis, *Annals of Mathematical Statistics*, **34**, 122–148.

Anderson TW and Rubin H (1956). Statistical inference in factor analysis. In *Proceedings of 3rd Berkeley Symposium on Mathematical Statistics and Probability*, **5**, University of California Press, 111–150.

Chen J, Li P, and Fu Y (2012). Inference on the order of a normal mixture, *Journal of the American Statistical Association*, **107**, 1096–1105.

Cook RD (1994). Using dimension-reduction subspaces to identify important inputs in models of physical systems. In *Proceedings of the Section on Physical and Engineering Sciences* (pp. 18–25), American Statistical Association, Alexandria, VA.

Cook RD (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*, John Wiley & Sons, New York.

Cook RD and Weisberg S (1991). Comment on sliced inverse regression by K. C. Li, *Journal of the American Statistical Association*, **86**, 328–332.

Cook RD and Weisberg S (1994). *An Introduction to Regression Graphics*, John Wiley & Sons, New York.

Gentle JE (2007). *Matrix Algebra: Theory, Computations, and Applications in Statistics*, Springer, New York.

Jeffries NO (2003). A note on 'Testing the number of components in a normal mixture', *Biometrika*, **90**, 991–994.

Jolliffe IT (2002). *Principal Component Analysis*(2nd ed), Springer, New York.

Lawley DN (1953). A modified method of estimation in factor analysis and some large sample results. In *Uppsala Symposium on Psychological Factor Analysis* (pp. 35-42), Munksgaards, Copenhagen.

Li B, Zha H, and Chiaromonte F (2005). Contour regression: a general approach to dimension reduction, *Annals of Statistics*, **33**, 1580–616.

Li KC (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma, *Journal of the American Statistical Association*, **87**, 1025–1039.

Li KC (1991). Sliced inverse regression for dimension reduction, *Journal of the American Statistical Association*, **86**, 316–327.

Lo Y, Mendell NR, and Rubin DB (2001). Testing the number of components in a normal mixture, *Biometrika*, **88**, 767–778.

Lo Y (2005). Likelihood ratio tests of the number of components in a normal mixture with unequal variances, *Statistics & Probability Letters*, **71**, 225–235.

Meyer CD (2000). *Matrix Analysis and Applied Linear Algebra*, Society for Industrial and Applied Mathematics, Philadelphia, PA.

Paisley J and Carin L (2009). Nonparametric factor analysis with beta process priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*(pp. 777–784), ACM, New York.

Scrucca L (2011). Model-based SIR for dimension reduction, *Computational Statistics & Data Analysis*, **55**, 3010–3026.

Seo B and Kim D (2012). Root selection in normal mixture models, *Computational Statistics & Data Analysis*, **56**, 2454–2470.

Tipping ME and Bishop CM (1999a). Probabilistic principal component analysis, *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, **61**, 611–622.

Tipping ME and Bishop CM (1999b). Mixtures of probabilistic principal component analyzers, *Neural Computation*, **11**, 443–482.

Vidal R (2011). Subspace clustering, *IEEE Signal Processing Magazine*, **28**, 52–68.

Whittle P (1952). On principal components and least square methods of factor analysis, *Scandinavian Actuarial Journal*, **36**, 223–239.

Young G (1941). Maximum likelihood estimation and factor analysis, *Psychometrika*, **6**, 49–53.

Received November 14, 2016; Revised December 12, 2016; Accepted December 13, 2016