International Journal of Statistical Sciences *ISSN 1683–5603*
Vol. 6 (Special Issue), 2007, pp 351-361

Mixture Models for Exploring Local PCA Structures

M. Nurul Haque Mollah
Dept. of Statistics
University of Rajshahi
Rajshahi-6205
Bangladesh
E-mail: mnhmollah@yahoo.co.in

[Received December 11, 2005; Revised June 23, 2007; Accepted September 5, 2007]

Abstract

Principal component analysis (PCA) is one of the most popular technique for dimensionality reduction of multivariate data. This paper discusses a new learning algorithm to explore local PCA structure in which the observed data follow a mixture of several PCA models, where each model is described by a linear combination of independent and Gaussian sources. The proposed method is based on a mixture of several Gaussian distributions to extract all local PCA structures simultaneously. Parameters are estimated by maximizing likelihood function. The performance of the proposed method is compared with some existing PCA algorithms using synthetic datasets.

Keywords and Phrases: Principal Component Analysis (PCA), PCA mixture models, Local PCA structure, Gaussian mixture distribution, Maximum likelihood estimation.

AMS Classification: 62H25, 35B50.

1 Introduction

Principal component analysis (PCA) is one of the most popular technique for processing, compressing and visualizing multivariate data. It is widely used for reducing dimensionality of multivariate data (1). In general, PCA aims to extract the most informative q-dimensional output vector $\boldsymbol{y}(t)$ from input vector $\boldsymbol{x}(t) = \Big(x_1(t), x_2(t), ..., x_m(t)\Big)^T$ of dimension $m \geq q$ whose components are assumed to be

Gaussian and linearly correlated with each other. This is achieved by learning the $m \times q$ orthogonal matrix Γ or $\Gamma^T\Gamma = I_q$ (identity matrix) which relates $\boldsymbol{x}(t) \sim N(\boldsymbol{\mu}, \Sigma)$ to $\boldsymbol{y}(t) = \Big(y_1(t), y_2(t), ..., y_q(t)\Big)^T$ by

$$\boldsymbol{y}(t) = \widehat{\Gamma}^T\Big(\boldsymbol{x}(t) - \widehat{\boldsymbol{\mu}}\Big), \quad t = 1, 2, \ldots, n \tag{1}$$

such that components of $\boldsymbol{y}(t)$ are mutually uncorrelated satisfying $\mathrm{Var}(Y_1) > \mathrm{Var}(Y_2) > ... > \mathrm{Var}(Y_q) > 0$, where $\widehat{\Gamma}$ is obtained as the q dominant eigenvectors of the estimated covariance matrix $\widehat{\Sigma}$, which we write in the form

$$\widehat{\Gamma} = \mathrm{eigen}(\widehat{\Sigma}). \tag{2}$$

Note that a dominant eigenvector of a covariance matrix is also known as principal subspace or principal component (PC) vector. The estimates $\Big(\widehat{\boldsymbol{\mu}}, \widehat{\Sigma}\Big)$ of $(\boldsymbol{\mu}, \Sigma)$ are obtained by maximizing likelihood function as given below:

$$\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{x}(t) \tag{3}$$

and

$$\widehat{\Sigma} = \frac{1}{n}\sum_{t=1}^{n} \Big(\boldsymbol{x}(t) - \widehat{\boldsymbol{\mu}}\Big)\Big(\boldsymbol{x}(t) - \widehat{\boldsymbol{\mu}}\Big)^T \tag{4}$$

The input vector $\boldsymbol{x}(t)$ is represented by the m-dimensional latent vector $\boldsymbol{s}(t) = \Big(s_1(t), s_2(t), ..., s_m(t)\Big)^T$ as

$$\boldsymbol{x}(t) = A\boldsymbol{s}(t) + \boldsymbol{b}, \quad t = 1, 2, \ldots, n \tag{5}$$

where A is an $m \times m$ coefficient matrix and $\boldsymbol{b}$ is a bias vector. The components of the latent vector $\boldsymbol{s}(t)$ are assumed to be mutually independent and Gaussian. The latent variable model expressed by equation 5 is considered as the data generating model. It offers a more economical explanation of the linear dependencies among the input observations (7; 8).

In standard PCA model defined by equations 1-5, all latent vectors belong to only one source class $\mathscr{S}$, and all input vectors belong to the same class in the entire data space $\mathscr{D}$. However, in practice, these source vectors may originate from several source classes, and the corresponding observed vectors belong to several classes in the entire data space. In this case, performance of the classical PCA methods are not so good. To overcome this problem, Tipping et al. (8) proposed mixture of probabilistic

PCA. However, their method is a little bit inconvenient due to computational and conceptual complexity. Therefore, an attempt is made to propose a new PCA mixture model for exploring local PCA structures. Note that in the PCA mixture model, the observed data in each class are considered to be a linear combination of independent and Gaussian sources (8; 5). When the data in each class are modeled as multivariate non-Gaussian, it is known as a ICA mixture model (2; 4).
Section 2 discusses a new PCA mixture model, section (2.1) describes the derivation of the PCA mixture model. In section (2.2), we discuss how to extract local PCA structures. Finally, section (3) presents numerical examples, and Section (4) presents the conclusions of this study.

2 The Proposed PCA Mixture Model

Let us assume that source vectors come from c source classes $\{\mathscr{S}_1, \mathscr{S}_2, ..., \mathscr{S}_c\}$ and that the corresponding observed vectors belong to c different data classes $\{\mathscr{D}_1, \mathscr{D}_2, ..., \mathscr{D}_c\}$ in the entire data space $\mathscr{D}$, where c is assumed to be known in advance. In addition, we assume that data class $\mathscr{D}_k$ occurs in the entire data space $\mathscr{D}$ due to the source class $\mathscr{S}_k$, ($k = 1, 2, \ldots, c$). In practice, the occurrence order of an observed vector in the entire data space $\mathscr{D}$ from a source class is unknown. However, we can assume that an observed vector $\boldsymbol{z}_k(j) \in \mathscr{D}_k = \{\boldsymbol{z}_k(j);\ j = 1, 2, \ldots, n_k\}$, $(k = 1, 2, \ldots, c;\ \sum_{k=1}^{c} n_k = n)$ whose occurrence order is unobserved, follows a PCA data generating model expressed as

$$\boldsymbol{z}_k(j) = A_k \boldsymbol{s}_k(j) + \boldsymbol{b}_k, \tag{6}$$

where A_k is an $m \times m$ mixing matrix, $\boldsymbol{b}_k$ is the bias vector and $\boldsymbol{s}_k(j) \in \mathscr{S}_k = \{\boldsymbol{s}_k(j);\ j = 1, 2, \ldots, n_k\}$, $(k = 1, 2, \ldots, c)$ is the j-th random vector in the source class k whose components are assumed to be independent and Gaussian. In a practical situation, an observable vector $\boldsymbol{x}(t) \in \mathscr{D} = \{\boldsymbol{x}(t);\ t = 1, 2, \ldots, n\}$ is obtained as a vector of $\bigcup_{k=1}^{c} \mathscr{D}_k = \{\boldsymbol{z}_k(j);\ j = 1, 2, \ldots, n_k,\ k = 1, 2, \ldots, c;\ \sum_{k=1}^{c} n_k = n\}$ such that $\mathscr{D} = \bigcup_{k=1}^{c} \mathscr{D}_k$. If the permutation of $\{\boldsymbol{z}_1(1), \boldsymbol{z}_1(2), \ldots, \boldsymbol{z}_k(j), \ldots, \boldsymbol{z}_c(n_c)\}$ into $\{\boldsymbol{x}(1), \boldsymbol{x}(2), \ldots, \boldsymbol{x}(n)\}$ is purely random, then equation 6 reduces to the PCA mixture model and an observed random vector $\boldsymbol{x}(t)$ follows Gaussian mixture distribution (3) as

$$p(\boldsymbol{x}(t) \mid \Theta) = \sum_{k=1}^{c} p(C_k) \varphi(\boldsymbol{x}(t) \mid \theta_k, C_k), \tag{7}$$

where $\Theta = \{\theta_1, ...\theta_k, ...\theta_c\}$, $\sum_{k=1}^{c} p(C_k) = 1$ and $\theta_k = \{\boldsymbol{\mu}_k, \Sigma_k\}$ are the unknown parameters for the Gaussian density

$$\varphi(\boldsymbol{x} \mid \theta_k, C_k) = |\det(2\pi\Sigma_k)|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k)\right\}, \tag{8}$$

corresponding to the data class C_k. The task is to estimate $(\boldsymbol{\mu}_k, \Sigma_k)$ for (1) and (2).

2.1 Derivation of the PCA Mixture Model by Maximizing Likelihood Function

Let $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n$ be a random sample drawn from (7), then the log-likelihood of the data for the unknown parameter $\Theta = \{\theta_1, \theta_2, ...\theta_c\}$ is given by

$$L = \sum_{t=1}^{n} \log p(\boldsymbol{x}(t) \mid \Theta). \tag{9}$$

The gradients of the parameters for class k is given by

$$\frac{\partial L}{\partial \theta_k} = \sum_{t=1}^{n} \frac{1}{p(\boldsymbol{x}(t) \mid \Theta)} \frac{\partial}{\partial \theta_k} p(\boldsymbol{x}(t) \mid \Theta) \tag{10}$$

$$= \sum_{t=1}^{n} \frac{\frac{\partial}{\partial \theta_k} p(C_k)\varphi(\boldsymbol{x}(t) \mid \theta_k, C_k)}{p(\boldsymbol{x}(t) \mid \Theta)} \tag{11}$$

Using the Bayes relation, the class probability for a given data vector $\boldsymbol{x}(t)$ is

$$p(C_k \mid \boldsymbol{x}(t), \Theta) = \frac{p(C_k)\varphi(\boldsymbol{x}(t) \mid \theta_k, C_k)}{\sum_{k=1}^{c} p(C_k)\varphi(\boldsymbol{x}(t) \mid \theta_k, C_k)} \tag{12}$$

Substituting (12) in (11) leads to

$$\frac{\partial L}{\partial \theta_k} = \sum_{t=1}^{n} \frac{p(C_k \mid \boldsymbol{x}(t), \Theta)}{p(C_k)\varphi(\boldsymbol{x}(t) \mid \theta_k, C_k)} \frac{\partial}{\partial \theta_k} p(C_k)\varphi(\boldsymbol{x}(t) \mid \theta_k, C_k) \tag{13}$$

$$= \sum_{t=1}^{n} p(C_k \mid \boldsymbol{x}(t), \Theta) \frac{\partial}{\partial \theta_k} \log \varphi(\boldsymbol{x}(t) \mid \theta_k, C_k) \tag{14}$$

Now,

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} \log \varphi(\boldsymbol{x}(t) \mid \theta_k, C_k) = \Sigma_k^{-1}(\boldsymbol{x}(t) - \boldsymbol{\mu}_k)$$

and

$$\frac{\partial}{\partial \Sigma_k} \log \varphi(\boldsymbol{x}(t) \mid \theta_k, C_k) = \frac{1}{2}\Sigma_k^{-1}\{(\boldsymbol{x}(t) - \boldsymbol{\mu}_k)(\boldsymbol{x}(t) - \boldsymbol{\mu}_k)^T - \Sigma_k\}\Sigma_k^{-1}$$

Therefore, $\frac{\partial L}{\partial \boldsymbol{\mu}_k} = 0$ implies

$$\boldsymbol{\mu}_k{}^* = \frac{\sum_{t=1}^{n} p(C_k \mid \boldsymbol{x}(t), \Theta)\boldsymbol{x}(t)}{\sum_{t=1}^{n} p(C_k \mid \boldsymbol{x}(t), \Theta)} \tag{15}$$

and $\frac{\partial L}{\partial \Sigma_k} = 0$ implies

$$\Sigma_k{}^* = \frac{\sum_{t=1}^{n} p(C_k \mid \boldsymbol{x}(t), \Theta)(\boldsymbol{x}(t) - \boldsymbol{\mu}_k)(\boldsymbol{x}(t) - \boldsymbol{\mu}_k)^T}{\sum_{t=1}^{n} p(C_k \mid \boldsymbol{x}(t), \Theta)} \tag{16}$$

Note that prior probability $p(C_k)$ can be updated by

$$p(C_k)^* = \frac{1}{n} \sum_{t=1}^{n} p(C_k \mid \boldsymbol{x}(t), \Theta) \tag{17}$$

The notations $\boldsymbol{\mu}_k{}^*$, $\Sigma_k{}^*$ and $p(C_k)^*$ are the update of $\boldsymbol{\mu}_k$, Σ_k and $p(C_k)$ respectively, where Σ_k should be initialized by identity matrix and other parameters can be initialized randomly. Then, the orthogonal matrix for extracting k-th local PCA structure is obtained as

$$\widehat{\Gamma}_k = \text{eigen}\left(\widehat{\Sigma}_k\right) \tag{18}$$

If c=1, then the proposed PCA algorithm reduces to the standard PCA algorithm as discussed around equations 1-5.

2.2 How to Extract Local PCA Structures

For local PCA, we transform the input data vector $\boldsymbol{x}(t)$ into output vector $\boldsymbol{y}(t)$ by

$$\boldsymbol{y}(t) = \widehat{\Gamma}_{(k)}^T \left(\boldsymbol{x}(t) - \widehat{\boldsymbol{\mu}}_{(k)}\right), \quad t = 1, 2, ..., n; \;\; k = 1, 2, ..., c \tag{19}$$

where

$$\left(\widehat{\boldsymbol{\mu}}_{(k)}, \widehat{\Gamma}_{(k)}\right) \in \left\{ \left(\widehat{\boldsymbol{\mu}}_1, \widehat{\Gamma}_1\right), \left(\widehat{\boldsymbol{\mu}}_2, \widehat{\Gamma}_2\right), ..., \left(\widehat{\boldsymbol{\mu}}_c, \widehat{\Gamma}_c\right) \right\}, \quad k = 1, 2, ..., c$$

Then, (k)-th local PCA structure is defined by those output vectors $\boldsymbol{y}(t)$ whose input vectors $\boldsymbol{x}(t)$ belong to the data class

$$\mathscr{D}_{(k)} = \left\{ \boldsymbol{x}(t) \in \mathscr{D} : \; \boldsymbol{x}(t) = \underset{k \in \{1,2,...,c\}}{\text{argmax}} \; p(C_k \mid \boldsymbol{x}(t), \Theta) \right\} \tag{20}$$

3 Simulation and Discussion

To demonstrate the performance of the proposed algorithm, we generated the following data sets from the c-component Gaussian mixture distribution with different mean vectors $\boldsymbol{\mu}_k$ and covariance matrices Σ_k, $(k = 1, 2, ..., c)$ using the data generating model (6).

Dataset 1 : 1000 samples were generated from two components Gaussian mixture distribution (Figure 14a) with mean vectors $\boldsymbol{\mu}_1 = (0,0)^T$ and $\boldsymbol{\mu}_2 = (8,0)^T$, and variance matrices

$$\Sigma_1 = \begin{bmatrix} 3.5 & 2.1 \\ 2.1 & 2.4 \end{bmatrix} \text{ and } \Sigma_2 = \begin{bmatrix} 3.6 & -1.2 \\ -1.2 & 1.8 \end{bmatrix}$$

respectively, where the first 500 source random samples were drawn from $N(\boldsymbol{\mu}_1, \Sigma_1)$ and the rest 500 source random samples were drawn from $N(\boldsymbol{\mu}_2, \Sigma_2)$.

Dataset 2 :1000 samples were generated from two components Gaussian mixture distribution (Figure 15a) with mean vectors $\boldsymbol{\mu}_1 = (0,0,0,0,0)^T$ and $\boldsymbol{\mu}_2 = (8,5,1,0,0)^T$, and variance matrices

$$\Sigma_1 = \begin{bmatrix} 8.3 & 0 & 0 & 0 & 0 \\ 0 & 6.7 & 0 & 0 & 0 \\ 0 & 0 & 0.8 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0.2 \end{bmatrix} \text{ and } \Sigma_2 = \begin{bmatrix} 1.5 & 0 & 0 & 0 & 0 \\ 0 & 0.9 & 0 & 0 & 0 \\ 0 & 0 & 0.6 & 0 & 0 \\ 0 & 0 & 0 & 0.2 & 0 \\ 0 & 0 & 0 & 0 & 0.1 \end{bmatrix}$$

respectively, where the first 500 source random samples were drawn from $N(\boldsymbol{\mu}_1, \Sigma_1)$ and the rest 500 source random samples were drawn from $N(\boldsymbol{\mu}_2, \Sigma_2)$.

For convenience of presentation, samples in both datasets are ordered by class. However, we will not use any information on sample order throughout the simulation study so that the simulation results must be the same even when samples are randomly ordered.

Let us first consider dataset 1 as early discussed which is consist of 2 data clusters,

where one cluster ($\mathscr{D}_1$) is represented by the symbol '.' and the other one ($\mathscr{D}_2$) by the symbol '$\times$' as shown in figure 14a. Figure 14b represent the scatter plot of 1st and 2nd PCs (principal components) obtained by standard method which contradicts with the uncorrelatedness properties of PCA. Therefore, standard PCA is not so good for this dataset.

By the proposed method, we obtain two sets of estimates $\{\hat{\boldsymbol{\mu}}_{(1)}, \hat{\Gamma}_{(1)}\}$ and $\{\hat{\boldsymbol{\mu}}_{(2)}, \hat{\Gamma}_{(2)}\}$, simultaneously. Figure 14c represent the scatter plot between 1st and 2nd PCs obtained by the estimates $\{\hat{\boldsymbol{\mu}}_{(1)}, \hat{\Gamma}_{(1)}\}$ using (19). It is seen that the output components of $\{\boldsymbol{y}(t)\}$ consist of two clusters, where one cluster with symbol '.' corresponding to data cluster $\mathscr{D}_1$ satisfies both uncorrelatedness and variance properties of PCA. Figure 14d shows the scatter plot between two PCs obtained by the estimates $\{\hat{\boldsymbol{\mu}}_{(2)}, \hat{\Gamma}_{(2)}\}$ using (19). The exhibited properties are the same as those found in Figure 15c and we obtain the second local PCA structure with the symbol '$\times$' corresponding to data cluster $\mathscr{D}_2$. Figures 14e and 14f show the class probability of each data point corresponding to the estimates $\{\hat{\boldsymbol{\mu}}_{(1)}, \hat{\Gamma}_{(1)}\}$ and $\{\hat{\boldsymbol{\mu}}_{(2)}, \hat{\Gamma}_{(2)}\}$, respectively. We see that in

each set of estimates, one data cluster was used and the other data cluster was totally ignored by the class probability. The arrows in Figures 14c and 14d represent the center of local PCA structures.

If there is only one data cluster in the entire data space (that is $c = 1$) and outlier does not exist, then standard PCA algorithm may be better than any other PCA algorithms. Therefore, we considered standard PCA results using only one data cluster to investigate the performance of the proposed method. First we considered data cluster $\mathscr{D}_1$ (Fig. 14g) from the entire data space $\mathscr{D}$ (Fig. 14a). Figure 14h represent the scatter plot of 1st and 2nd PCs obtained by standard method using data cluster $\mathscr{D}_1$. Comparing figures 14c and 14h, we see that standard PCA structure based on data cluster $\mathscr{D}_1$ and the local PCA structure corresponding to data cluster $\mathscr{D}_1$ are almost equivalent. Similarly, comparing figures 14d and 14j, we see that standard PCA structure based on data cluster $\mathscr{D}_2$ and the local PCA structure corresponding to data cluster $\mathscr{D}_2$ are also equivalent. Therefore, performance of the proposed method is good in our current context.

To investigate the performance of the proposed method in a comparisons of the mixture of PPCA algorithm, we consider dataset 2 as a multivariate Gaussian mixture data which consists of 2 clusters, where one cluster is represented by the symbol '.' and the other one by the symbol '$\times$'. With projection of observed data onto two-dimensional coordinates, two clusters are overlapped as shown in figure 15a. To compute the k-th local PCA structures, the true orthogonal matrix (Γ_k) will be identity matrix ($\mathbf{I}$) corresponding to k-th data cluster, since the covariance matrix corresponding to each cluster is diagonal for dataset 2. That is $\Gamma_k = (\gamma_{k1}, \gamma_{k2}, ..., \gamma_{kq}) = \mathbf{I}$. An estimate $\widehat{\Gamma}_k = (\widehat{\gamma}_{k1}, \widehat{\gamma}_{k2}, ..., \widehat{\gamma}_{kq})$ will be good for Γ_k if the inner product (IP) between γ_{ki} and $\hat{\gamma}_{kj}$ satisfy

$$\gamma_{ki}^T \hat{\gamma}_{kj} = \begin{cases} 1, & \text{for} \quad i = j \\ 0, & \text{for} \quad i \neq j \end{cases} \tag{21}$$

where $k = 1, 2, ..., c$. It should be noted here that the mixture of PPCA algorithm is able to estimate at most first $q = m - 1 = 5 - 1 = 4$ principle components (PCs) for five-dimensional dataset. Therefore we have to consider first 4 PCs obtained by the proposed method for comparison with the mixture of PPCA method. Both the proposed method and the mixture of PPCA method explore two orthogonal matrices $\widehat{\Gamma}_1 \& \widehat{\Gamma}_2$ for local PCA corresponding to each data cluster for dataset 2. However, in the case of standard PCA, $\widehat{\Gamma}_1 = \widehat{\Gamma}_2$.

Figures 15b and 15c show the inner products (IP) between true vector ($\gamma_{1i} \& \gamma_{2i}$) in $\Gamma_1 \& \Gamma_2$ and their estimates $\widehat{\gamma}_{1i} \& \widehat{\gamma}_{2i}$, ($i = 1, 2, .., 4$) in $\widehat{\Gamma}_1 \& \widehat{\Gamma}_2$, respectively. In each plot, dash-dot line with marker style (.), solid line with marker style ($\times$) and dotted line with marker style (o) represent the estimation based on standard PCA method, mixture of PPCA method and the proposed method, respectively. In both plots, clearly, we see that standard estimator does not satisfy $\gamma_{ki}^T \hat{\gamma}_{ki} = 1$ for $i = 1, 2$, while it is almost

satisfied for all $i = 1, 2, ..., 4$ with the estimates obtained by other two methods for $k = 1, 2$.
Now we compute the amount of error for the estimates $\hat{\gamma}_{ki}$ by the following formula

$$\text{Error}(ki) = \frac{1}{m} \left\| \gamma_{ki} - \hat{\gamma}_{ki} \times \text{sign}(\hat{\gamma}_{ki}(i)) \right\|^2, \tag{22}$$

where $\text{sign}(\hat{\gamma}_{ki}(i)) = \pm 1$, the sign of i-th component of $\hat{\gamma}_{ki}$. Figures 15d and 15e show the amount of error for the estimates $\hat{\gamma}_{ki}$, $(i = 1, 2, .., 4)$ for k =1 and 2, respectively. In each plot, dash-dot line with marker style (.), solid line with marker style ($\times$) and dotted line with marker style (o) represent the estimation based on standard PCA method, mixture of PPCA method and the proposed method as before, respectively. In both plots, clearly, we see that the amount of error for the standard estimates are large for $i = 1, 2$, while the amount of error for the estimates obtained by other two methods are almost close to zero for all $i = 1, 2, ..., 4$. Therefore, local PCA based on the mixtures of PPCA method and the proposed method are better than standard PCA method.

Figures 15f and 15g represent the percentage of total variation (TV) by the estimates $\hat{\gamma}_{ki}$, $(i = 1, 2, .., 4)$ for k =1 and 2, respectively. In each plot, dash-dot line with marker style (.), solid line with marker style ($\times$) and dotted line with marker style (o) represent the estimation based on standard PCA method, mixture of PPCA method and the proposed method as before, respectively. In both plots, we see that % of TV for each PC obtained by the mixtures of PPCA method and the proposed method both are almost similar, while the % of TV for each PC obtained by the standard method is not similar for first and second principal components with the other two methods.

Figures 15h and 15i represent the cumulative % of TV upto the estimates $\hat{\gamma}_{ki}$, $(i = 1, 2, .., 4)$ for k =1 and 2, respectively. In each plot, dash-dot line with marker style (.), solid line with marker style ($\times$) and dotted line with marker style (o) represent the estimation based on standard PCA method, mixture of PPCA method and the proposed method as before, respectively. In both plots, we see that cumulative % of TV for each PC obtained by the mixtures of PPCA method and the proposed method both are almost similar, while the cumulative % of TV for each PC obtained by the standard method is not similar for first and second principal components with the other two methods.

4 Conclusions

We proposed a learning algorithm for exploring local PCA structure based on Gaussian mixture distribution. Parameters are estimated by maximizing the likelihood function. All local PCA structures can be extracted simultaneously from the entire data space using the proposed method.

The purpose of the proposed method is similar to the mixture of PPCA method proposed by Tipping and Bishop (8). The proposed procedure is able to find all principal components as well as minor components, whereas, mixture of PPCA method is not able to find some minor components.
Finally, we compare the performance of the proposed method with the standard PCA algorithm and the mixture of PPCA algorithm by simulation study. We observed the performance of the standard PCA method is not good in our current context, however, the performance of both the proposed method and the mixture of PPCA method is good and almost similar in the current context.
The main advantage of the proposed method is that it is more straightforward computationally and conceptually than mixture of PPCA method. However, either the proposed method or the mixture of PPCA method are not robust against outliers. A robust version of PCA mixture model is under development by Mollah et al. (6) and will be made available soon.

References

[1] Jolliffe, I. T., (2002): *Principal Component Analysis*, Springer.

[2] Lee, T.-W., Lewicki, M. S. and Sejnowski, T. J. (2000): ICA Mixture Models for unsupervised classification of non-Gaussian classes and automatic context switching in blind signal separation. *IEEE Trans. on Pattern Analysis an Machine Int.*, 22, pp. 1078-1089.

[3] McLachlan, G. J., and Peel, D. (2000): *Finite Mixture Models*, New York, Wiley.

[4] Mollah, M.N.H., Minami, M. and Eguchi, S. (2006): Exploring Latent Structure of Mixture ICA models by minimum β-divergence method. *Neural Computation* **18**(1), pp. 166-190.

[5] Mollah, M.N.H., Sultana, N., Minami, M. and Eguchi, S. (2005): Exploring Local PCA Structure for Dimensionality Reduction by Minimizing β-Divergence, *Research Memorandum No. 956*, The Institute of Statistical Mathematics, Tokyo.

[6] Mollah, M.N.H., Sultana, N., Minami, M. and Eguchi, S. (2007): Robust Extraction of Local Structures by the Minimum β-Divergence Method. (Submitted for publication)

[7] Tipping, M.E. and Bishop, C.M. (1997): Probabilistic principal component analysis. *J. Royal Statistical Society B, 61, Part 3, pp. 611-622*

[8] Tipping, M.E. and Bishop, C.M. (1999): Mixtures of Probabilistic principal component analysers. *Neural Computation* **11(2)**, pp. 443-482.

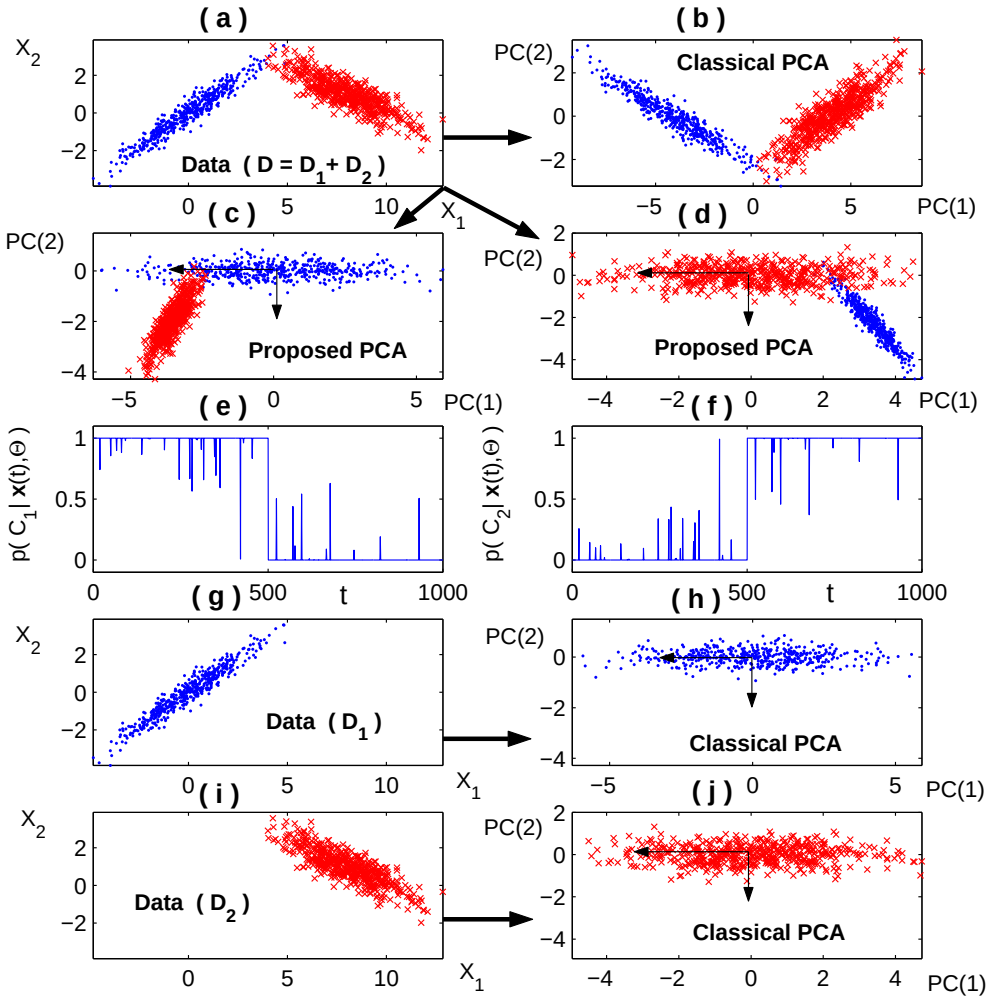

Figure 14: (a) Scatter plot of observed components belonging to the entire data space $\mathscr{D}$. (b) Scatter plot between first and second PCs estimated by standard method. (c-d) Scatter plot between first and second PCs estimated by the proposed method. (e-f) Class probability for each data point obtained by the proposed method. (g) Scatter plot of observed components belonging to data cluster $\mathscr{D}_1$. (h) Scatter plot between first and second PCs obtained by standard method using data cluster $\mathscr{D}_1$. (i) Scatter plot of observed components belonging to data cluster $\mathscr{D}_2$. (j) Scatter plot between first and second PCs obtained by standard method using data cluster $\mathscr{D}_2$.

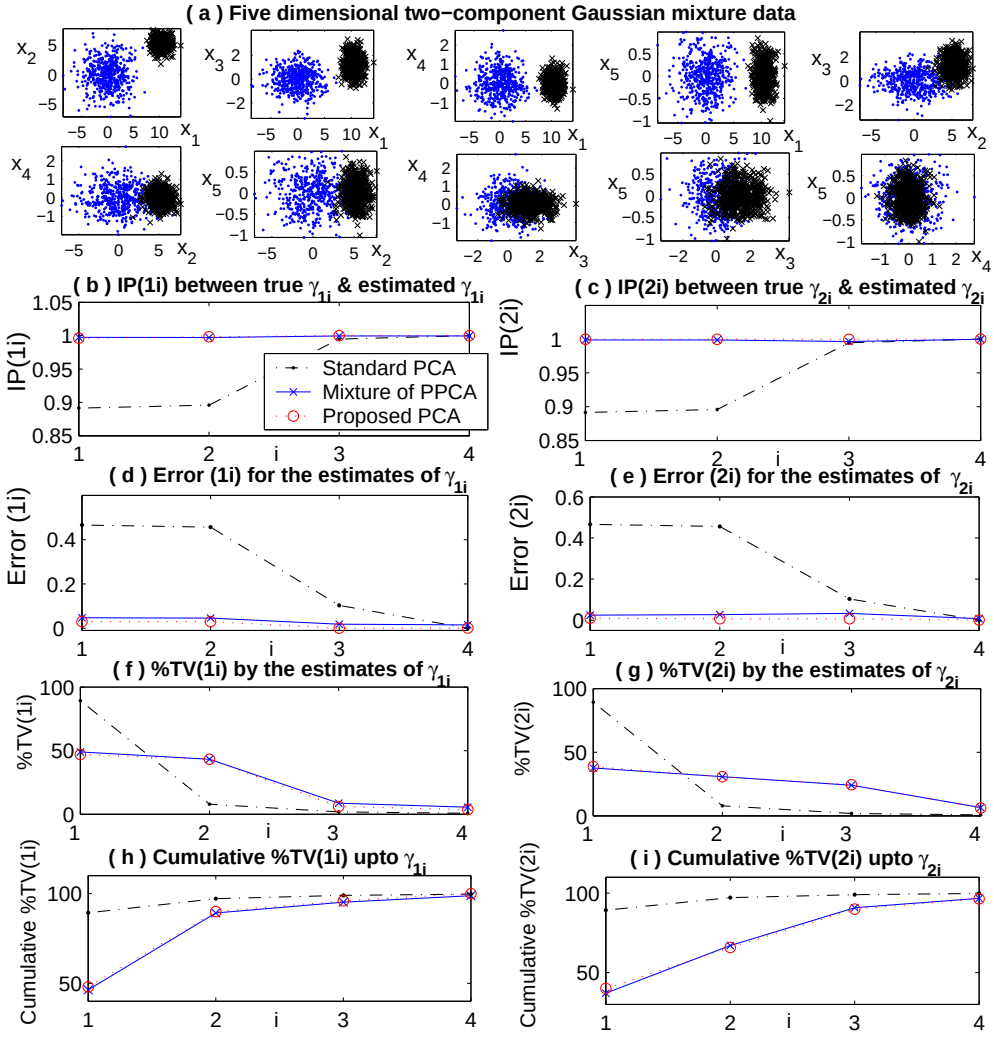

Figure 15: (a) Five dimensional two-component Gaussian mixture data displayed with 2×2 windows. (b-c) Inner product (IP) between true column vector in Γ_1&Γ_2 and the estimated column vector in $\widehat{\Gamma}_1$&$\widehat{\Gamma}_2$, respectively. (d-e) Amount of error for the estimates $\widehat{\gamma}_{1i}$&$\widehat{\gamma}_{2i}$, (i=1,2,3,4) in $\widehat{\Gamma}_1$&$\widehat{\Gamma}_2$, respectively. (f-g) Percentage of total variation (%TV) by $\widehat{\gamma}_{1i}$&$\widehat{\gamma}_{2i}$, (i=1,2,3,4) in $\widehat{\Gamma}_1$&$\widehat{\Gamma}_2$, respectively. (h-i) Cumulative %TV upto $\widehat{\gamma}_{1i}$&$\widehat{\gamma}_{2i}$, (i=1,2,3,4) in $\widehat{\Gamma}_1$&$\widehat{\Gamma}_2$, respectively.