An Effective EM Algorithm for PCA Mixture Model

Zhong Jin[1,3], Franck Davoine[2], and Zhen Lou[3]

[1] Centre de Visio per Computador
Universitat Autonoma de Barcelona
08193 Barcelona, Spain
zhong.jin@cvc.uab.es
[2] Heudiasyc - CNRS Mixed Research Unit
Compiegne University of Technology
60205 Compiegne cedex, France
franck.davoine@hds.utc.fr
[3] Department of Computer Science
Nanjing University of Science and Technology
Nanjing, People's Republic of China

Abstract. This paper studied PCA mixture model in high dimensional space. A novel EM learning approach by using perturbation was proposed for the PCA mixture model. Experiments showed the novel perturbation EM algorithm is more effective in learning PCA mixture model than an existing constrained EM algorithm.

1 Introduction

In recent years, there has been increasing interest in PCA mixture model. Mixture model provides a simple framework for modelling data complexity by a weighted combination of component distributions [1, 2]. It has been widely used in machine learning, image processing, and data mining due to its great flexibility and power. However, since the component distributions in mixture model are usually formalized to be probability density functions, there are limited applications to practical problems in high dimensional space.

As a variation of mixture model, PCA mixture model is proposed to use principal component analysis (PCA) to express component distributions. The idea of PCA mixture model is motivated by a mixture-of-experts technique that models a non-linear distribution by a combination of local linear sub-models, each with a relatively simple distribution [3, 4]. Hinton et al. [5] proposed a PCA mixture model based on the reconstruction error. Tipping and Bishop defined a mixture model for probabilistic principal component analyzers (PPCA), whose parameters can be determined using an EM algorithm [6]. Recently, mixtures of probabilistic principal component analyzers was used to model data that lies on or near a low dimensional manifold in a high dimensional observation space [7]. Kim et al. discussed the problem to select model order for PCA mixture model [8].

A. Fred et al. (Eds.): SSPR&SPR 2004, LNCS 3138, pp. 626–634, 2004.

Expectation-maximization (EM) algorithm is a powerful algorithms for maximum likelihood (ML) or maximum a posterior (MAP) estimation in problems with incomplete data, e.g., fitting a mixture model to observed data [9, 10]. The EM algorithm provides iterative formulae for the estimation of the unknown parameters of the mixture. The drawbacks of EM algorithm for a problem in high dimensional space is that the mixing components are assumed to have probability density functions in the data space. In practical problems, the mixing components may only have probability density functions in low dimensional manifolds.

In this paper, the EM algorithm for PCA Gaussian mixture model is studied. It is organized as follows. Section 2 gives an introduction to the Gaussian mixture model. Section 3 proposes a new EM algorithm for PCA mixture model. Experiments are performed in Section 4. Finally, conclusions are drawn in Section 5.

2 The EM Algorithm for Gaussian Mixtures

A Gaussian mixture is defined as a combination of Gaussian densities. A Gaussian density in a d-dimensional space, parameterized by its mean $\mathbf{m} \in \Re^{\mathbf{d}}$ and $d \times d$ covariance matrix C, is defined by the density:

$$\phi(x;\theta) = \frac{1}{(2\pi)^{\frac{d}{2}}|C|^{\frac{1}{2}}} exp\{-\frac{(x-m)^t C^{-1}(x-m)}{2}\}, \tag{1}$$

where $\theta = (m, C)$.

A mixture of k Gaussian densities is then defined as:

$$f_k(x) = \sum_{j=1}^{k} \pi_j \phi(x;\theta_j), \tag{2}$$

with

$$\sum_{j=1}^{k} \pi_j = 1, \tag{3}$$

where $\theta_j = (m_j, C_j)$ and $\pi_j \geq 0$, for $j = 1, 2, \cdots, k$. The π_j are called the mixing weights and $\phi(x;\theta_j)$ the components of the mixture.

A training set $X_n = \{x_1, x_2, \ldots, x_n\}$ of independent and identically distributed points $x_i \in \Re^d$ is assumed to be sampled from Eq. (2). The task is to estimate the parameters of the mixture that maximize the log-likelihood

$$\mathbf{L} = \frac{1}{n} \sum_{i=1}^{n} \log f_k(x_i). \tag{4}$$

The Expectation-Maximization (EM) algorithm is a well-known statistical tool for maximum likelihood problems involving hidden or unobserved variables [9]. In the case of mixtures, the unobservable variable can be regarded

as the component from which each input point has been sampled. The EM algorithm enables us to update the parameters of a given k-component mixture with respect to X_n such that the log-likelihood of X_n is never smaller under the new mixture.

Let

$$r_{ji} = P(j|x_i) \tag{5}$$

be the posterior probability that the point x_i is sampled from the ^{j}th mixing component. The EM iteration consists of one expectation step (E-step) and one maximization step (M-step) as follows.

- **E-step**
  By the Bayes rule, the expectation values of r_{ji} can be given from the mixture model of Eq. (2) from the previous iteration:

$$r_{ji} = \frac{\pi_j \phi(x_i; \theta_j)}{f_k(x_i)}, \tag{6}$$

  where $\theta_j = (m_j, C_j)$.
- **M-step**
  The component parameters can be estimated from samples and the expectation values of r_{ji} of Eq. (6):

$$\pi_j = \frac{1}{n} \sum_{i=1}^{n} r_{ji}, \tag{7}$$

$$m_j = \frac{1}{n\pi_j} \sum_{i=1}^{n} r_{ji} x_i, \tag{8}$$

$$C_j = \frac{1}{n\pi_j} \sum_{i=1}^{n} r_{ji} (x_i - m_j)(x_i - m_j)^t. \tag{9}$$

3 EM Algorithm for PCA Mixture Model

In this section, we will discuss the limitation of EM in high dimensional space firstly and then propose a novel EM algorithm for PCA mixture model.

3.1 Limitation of EM

In a high dimensional space, some mixing components may lies on or near a low dimensional manifold. In other words, for some mixing components, the covariance matrices C may be singular in high dimensional space so that there are not Gaussian densities of Eq. (1). In this case, Eq. (1), then E-step of Eq. (6), can not be conducted since there are not inverse matrices for such mixing components which are degenerate in a low dimensional manifold.

3.2 PCA Mixture Model

One solution to the above limitation of EM is to describe Gaussian density of Eq. (1) by using principal component analysis (PCA).

Let $\Psi = (\psi_1, \cdots, \psi_d)$ be the matrix whose columns are the unit-norm eigenvectors of the covariance matrix C of Eq. (1). Let $\Lambda = diag(\lambda_1, \cdots, \lambda_d)$ be the diagonal matrix of the eigenvalues of C, where λ_i are the eigenvalues corresponding to the eigenvectors ψ_i $(i = 1, \cdots, d)$. We have

$$\Psi^t C \Psi = \Lambda. \tag{10}$$

Let

$$y = \Psi^t (x - m), \tag{11}$$

where $y = [y(1), y(2), \cdots, y(d)]^{'}$ and $y(i)$ is the ^{i}th principal component of the d-dimensional vector y.

Assume that λ_i $(i = 1, \cdots, d)$ are ranked in order from larger to smaller as follows:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0. \tag{12}$$

The covariance matrix C can be any covariance matrix of the k mixing components and the mean m can be any mean of the k mixing components. Let C be C_j, the covariance matrix of the ^{j}th mixing component. If C_j is non-singular, we will have $\lambda_d > 0$. Then, the Gaussian density of Eq. (1) in x space can be expressed in y space as follows:

$$\phi(y; \theta_j) = \prod_{i=1}^{d} \frac{1}{(2\pi\lambda_i)^{\frac{1}{2}}} exp\{-\frac{[y(i)]^2}{2\lambda_i}\}. \tag{13}$$

Note that y not only depends on x, but also depends on $\theta_j = (m_j, C_j)$.

If C_j is singular, some last λ's will have a value of zero. This means that the ^{j}th mixing component is distributed in a low dimensional manifold in original x space. So, there doesn't exist a Gaussian density of Eq. (13) in x space for the mixing component.

Assume that

$$\lambda_1 \geq \cdots \geq \lambda_{d_j} > \lambda_{d_{j+1}} = \cdots = \lambda_d = 0, \tag{14}$$

where d_j is the number of non-zero eigen-values of the covariance matrix C_j of the ^{j}th mixing component. For the ^{j}th mixing component, there exists a Gaussian density in a low dimensional $[y(1), y(2), \cdots, y(d_j)]$ space as follows:

$$\phi(y; \theta_j) \approx \prod_{i=1}^{d_j} \frac{1}{(2\pi\lambda_i)^{\frac{1}{2}}} exp\{-\frac{[y(i)]^2}{2\lambda_i}\}. \tag{15}$$

The $d_j (j = 1, \cdots, k)$ in Eq. (15) can be called as an order of principal components for the ^{j}th mixing component. If d_j is equal to d, i.e., $d_j = d$, Eq. (15) is the same as Eq. (13). In other word, Eq. (13) is a special case of Eq. (15).

What is a PCA mixture model? Now, we can give a definition. We define the mixture model of Eqs. (2), (3) and (15) as a PCA mixture model, where $d_j(j = 1, \cdots, k)$, the orders of principal components for the mixing components, can be determined by Eq. (14) or by any other assumptions. This definition can be regarded as a generalization of the PCA mixture model used in [8], where $d_j(j = 1, \cdots, k)$ is assumed to be the same value.

Note that the $f_k(x)$ in the PCA mixture model is not a probability density function now. It is just a structure description of the data set in high dimensional space.

3.3 A Perturbation Approach

Recently, Kim et al. discussed the problem to select model order for PCA mixture model [8], and proposed a constraint with $d_j(j = 1, \cdots, k)$ as follows:

$$d_1 = d_2 = \cdots = d_k. \tag{16}$$

Under the constraint of Eq. (16), $d_j(j = 1, \cdots, k)$ have to be the smallest of all the orders of principal components for the mixing components. This assumption increases the difficulty in learning the PCA mixture model.

In general, when $d_j < d$, no matter whether $d_j(j = 1, \cdots, k)$ are equal to each other, it is difficult to estimate r_{ji} in classical E-step by using Eq. (15) directly. The reason is that Eq. (15) may be a probability density in a different space for the different $j(j = 1, \ldots, k)$, even under the constraint of Eq. (16).

Our idea is to assign a very small positive value ε to the last eigenvalues $\lambda_{d_{j+1}}, \cdots, \lambda_d$ in Eq. (14). So, we have

$$\lambda_1 \geq \cdots \geq \lambda_{d_j} > \lambda_{d_{j+1}} = \cdots = \lambda_d = \varepsilon > 0, \tag{17}$$

where ε can be called a perturbation factor.

By introducing a perturbation factor in Eq. (17), we make the covariance matrix C_j of the ^{j}th mixing component be non-singular. Therefore, Eq. (13) can be directly used to estimate the raw values of r_{ji} of Eq. (5) as follows:

$$r_{ji} = \pi_j \phi(y; \theta_j), \tag{18}$$

where according to Eq. (11), $y = y_i$ is obtained from the point x_i instead of x, and $\theta_j = (m_j, C_j)$ instead of $\theta = (m, C)$.

Note that all the raw values $r_{ji}(j = 1, \cdots, k)$ will further be normalized by

$$\sum_{j=1}^{k} r_{ji} = 1. \tag{19}$$

So, a novel E-step by using perturbation can be introduced as follows:

- **Perturbation E-Step**
  - ◇ For $C = C_j$, the covariance matrix of the ^{j}th mixing component, make a PCA decomposition of Eq. (10) , and obtain all the d eigenvalues λ_i $(i = 1, \cdots, d)$ and corresponding eigenvectors ψ_i $(i = 1, \cdots, d)$.
  - ◇ According to Eq. (17), determine d_j and assign ε to the last $d - d_j$ eigenvalues $\lambda_{d_{j+1}}, \cdots, \lambda_d$.
  - ◇ For each point $x_i (i = 1, \cdots, n)$, compute the projection of x_i on all the d principal components according to Eq. (11).
  - ◇ Update the expectation values of r_{ji} of Eq. (5) according to Eqs. (13,17, 18,19).

Our novel EM algorithm by using perturbation for PCA mixture model includes the novel perturbation E-step and the classical M-step of Eqs. (7,8,9).

4 Experiments

In this section, some experiments are performed to see if the novel EM algorithm by using perturbation is effective to learn both non-degenerate mixing component and degenerate mixing component.

4.1 Synthetic Data

Although PCA mixture model is motivated by difficulties in the estimating of distributions in high dimensional space, it is of some significance to have an initial evaluation on the learning algorithms with problems in low dimensional space.

For simplicity, a problem of only two mixing components in two dimensional space is considered: one with a degenerate density in one dimensional subspace, and the other with a non-degenerate density in two dimensional space. We have considered four examples as follows.

- **Example (a)** One mixing component has a degenerated density in the direction of $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$. The other has a covariance of $\begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix}$. The first principal directions of these two mixing components are orthogonal to each other.
- **Example (b)** One mixing component has a degenerated density in the direction of $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$. The other has a covariance of $\begin{bmatrix} 1 & 0.6 \\ 0.6 & 3 \end{bmatrix}$. The first principal directions of these two mixing components are approximately orthogonal to each other.
- **Example (c)** One mixing component has a degenerated density in the direction of $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$. The other has a covariance of $\begin{bmatrix} 1 & -0.6 \\ -0.6 & 3 \end{bmatrix}$. The first principal directions of these two mixing components are approximately parallel to each other.

- **Example (d)** One mixing component has a degenerated density in the direction of $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$. The other has a covariance of $\begin{bmatrix} 1 & -0.6 \\ -0.6 & 1 \end{bmatrix}$. The first principal directions of these two mixing components are parallel to each other.

We draw 100 samples respectively for each mixing component. These four examples are displayed in Fig. 1 respectively.

4.2 Experimental Results and Analysis

For each example, we have performed experiments to learn PCA mixture model by our perturbation EM algorithm and the constrained EM algorithm used in [8] with a constraint of Eq. (16).

The experimental results with four examples by our perturbation EM algorithm are displayed in Fig. 1 (a1), (b1), (c1) and (d1) respectively. The experimental results with four examples by the constrained EM algorithm are displayed in Fig. 1 (a2), (b2), (c2) and (d2) respectively. In Fig. 1, the distribution centers are marked in green color for all the learned mixing components, and the probability ellipses in blue color are drawn to enclose about 90% samples for all the learned mixing components if without degeneration.

From Fig. 1, we can have the following facts and discussion:

- Our perturbation EM algorithm is effective to learn PCA mixture model. For each example, the novel perturbation EM algorithm is effective to learn both non-degenerate mixing component and degenerate mixing component.
- The constrained EM algorithm [8] is not as effective as the perturbation EM algorithm to learn PCA mixture model. It becomes more difficult for the constrained EM algorithm to learn the mixing components when the first principal directions of two mixing components become more parallel to each other.

5 Conclusion and Future Work

In this paper, we have studied PCA mixture model in high dimensional space and have given an analysis to the limitation of EM algorithms for PCA mixture model. A novel EM learning approach by using perturbation is proposed for PCA mixture model. Experiments with synthetic data show the novel perturbation EM algorithm is more effective to learn both non-degenerate mixing component and degenerate mixing component in PCA mixture model than an existing constrained EM algorithm.

In our future work, we are focusing on investigating some learning algorithms for PCA mixture model for practical problems in high dimensional space, such as face recognition and facial expression analysis. In these problems, only a small number of samples are available, and a singular decomposition theorem is used to obtain some eigen-vectors for non-zero eigenvalues. In other words, for practical application problems, the dimension d may be too high to obtain all the d eigenvectors in Eq. (10) and the transformation from x-space to y-space in Eq. (11) is not available.

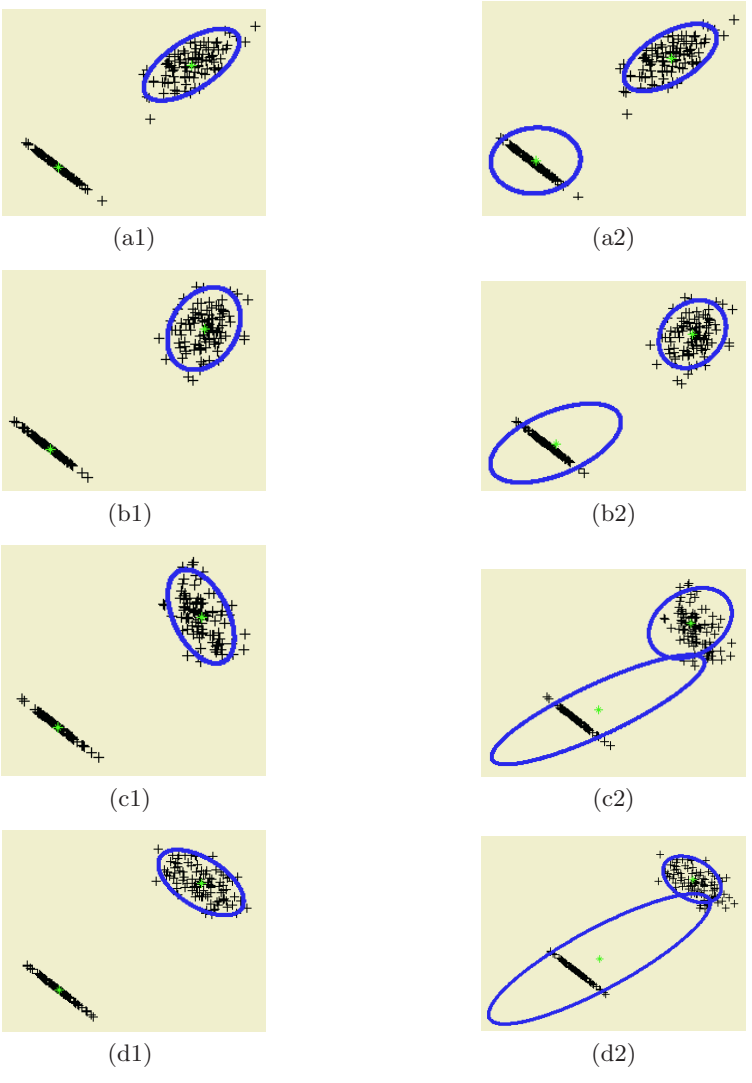

Fig. 1. Experiment results with four examples: (a1,b1,c1,d1) by our perturbation EM algorithm, and (a2,b2,c2,d2) by the constrained EM algorithm [8].

References

1. D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical analysis of finite mixture distribution*. Wiley, New York, 1985.
2. G. McLachlan and D. Peel. *Fixed mixture models*. Wiley, New York, 2000.
3. R. Jacobs, M. Jordan, S. Nowlan, and G. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

4. M. Jordan and R. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(5):181–214, 1994.
5. G. Hinton, P. Dayan, and M. Revow. Modeling the manifolds of images of handwritten digits. *IEEE Transactions on Neural Network*, 8(1):65–74, 1997.
6. M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analysers. *Neural Computation*, 11(2):443–482, 1999.
7. J.J. Verbeek, N. Vlassis, and B. Kröse. Coordinating mixtures of probabilistic principal component analyzers. Technical report, Computer Science Institute, University of Amsterdam, The Netherlands, Feb 2002. IAS-UVA-02-01.
8. Hyun-Chul Kim, Daijin kim, and Sung Yang Bang. An efficient model order selection for PCA mixture model. *Pattern Recognition Letters*, 24(9-10):1385–1393, 2003.
9. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39(1):1–38, 1977.
10. R. J. A. Little and D. B. Rubin. *Statistical analysis with missing data*. Wiley, New York, 1987.