Learning Probability Density Functions from Marginal Distributions with Applications to Gaussian Mixtures

Qutang Cai
Department of Automation
Tsinghua University, Beijing, China
caiqt98@mails.tsinghua.edu.cn

Changshui Zhang
Department of Automation
Tsinghua University, Beijing, China
zcs@mail.tsinghua.edu.cn

Chunyi Peng
Department of Automation
Tsinghua University, Beijing, China
pcy98@mails.tsinghua.edu.cn

Abstract— **Probability density function (PDF) estimation is a constantly important topic in the fields related to artificial intelligence and machine learning. This paper is dedicated to considering problems on the estimation of a density function simply from its marginal distributions. The possibility of the learning problem is first investigated and a uniqueness proposition involving a large family of distribution functions is proposed. The learning problem is then reformulated into an optimization task which is studied and applied to Gaussian mixture models (GMM) via the generalized expectation maximization procedure (GEM) and Monte Carlo method. Experimental results show that our approach for GMM, only using partial information of the coordinates of the samples, can obtain satisfactory performance, which in turn verifies the proposed reformulation and proposition.**

I. Introduction

Probability density function (PDF) estimation is very essential in artificial intelligence and machine learning, providing a solid basis for tasks such as probabilistic inference, clustering analysis, data mining and other related fields [1]. The PDF estimation methods can be divided into two categories: the parametric [2] and nonparametric approaches [3][4]. In the parametric approaches (e.g. maximum likelihood estimation (ML) and Bayesian estimation) the PDF is assumed to be of a given form and the function parameters are then optimized, while in the nonparametric approaches (e.g. histogram statistics, Parzen estimator and kernel density estimation) the form of the PDF need not to be known and the estimation is entirely driven by data. However, most of these contemporary parametric and nonparametric estimation techniques are applied in the original feature space, dealing with the full coordinate entries of samples. Even the classical methods for incomplete data such as Expectation Maximization method (EM) [5] still utilizes the full coordinates by introducing auxiliary latent variables. Till now, problems on learning the PDF simply from the given distributions in the subspaces of the original space have been seldom considered and these will be crucial when only marginal distributions can be observed(e.g. the computerized tomography). This paper investigates the availability of recovering the original PDF from the marginal distributions in given subspaces and finds it available for a wealth of density functions, and then proposes and studies an optimization task which maximizes the similarity with given marginal distributions. This task is then associated to the Gaussian Mixture Models (GMM), which have been found a number of important applications, such as video modelling [6], time series classification [7] and face representation [8], and the algorithm for learning GMM from marginal distributions are developed. The experimental results show that the proposed algorithm which uses only marginal distributions can achieve comparable performance as the classical algorithm [9] which uses the entire coordinate entries in the original space, and in turn verify the proposed proposition and optimization task.

The remainder of this paper is organized as follows. In section 2, we propose a uniqueness proposition for the learning problem and study an optimization task for the recovery. In section 3, the task is executed for the recovery of GMM, and the detailed algorithm based on the generalized EM and Monte Carlo method is developed. Experimental results and analysis are given in section 4, followed by some conclusions in section 5.

II. Main Properties and Problem Formulation

This section contains two parts: In the first part, the solvability of the learning problem is considered and the corresponding proposition is presented; In the second part, an optimization task for the learning purpose is studied and the conditions under which GMM can be learned via the optimization are proposed.

As is known, the marginal distributions always lose information of the original PDF, so one would imagine that some behaviors of the PDF may not emerge in all subspaces and it would happen that there are two or more PDFs which have the same marginal distributions. Thus the problem for learning general density functions merely from the marginal distributions seems unsolvable. However, the following proposition reveals that, for a very large family of density functions, they can be uniquely determined if all marginal distributions in one

0-7803-9048-2/05/$20.00 ©2005 IEEE

dimensional subspaces are known[1].

Proposition 1 (Uniqueness): Let $(x_1, x_2, \cdots, x_n)$ be the n-dimensional random variable in R^n with the probability density function $f(x_1, x_2, \cdots, x_n) \in L^2(R^n)$. Suppose for any given linear combination of $x_1, x_2, \cdots, x_n$, the marginal distribution can be known. Then the original distribution f can be uniquely determined.

Proof: Let $\widehat{f}(w_1, w_2, \cdots, w_n)$ be the characteristic function [2] for f, $a_1x_1 + a_2x_2 + \cdots + a_nx_n$ be an arbitrary linear combination of $x_1, x_2, \cdots, x_n$. By assumption, the distribution of $\sum_{i=1}^{n} a_ix_i$ is known, denoted by f_a. Then the characteristic function for f_a can be computed, written by $\widehat{f}_a(w)$. On the other hand,

$$\begin{aligned} \widehat{f}_a(w) &= \int f(x_1, \cdots, x_n) \exp(jw \sum_{i=1}^{n} a_ix_i) dx_1 \cdots dx_n \\ &= \widehat{f}(a_1w, a_2w, \cdots, a_nw). \end{aligned} \quad (1)$$

For any $(w_1, w_2, \cdots, w_n)$, take $a_i = w_i$ and $w = 1$ in (1), and hence $\widehat{f}(w_1, w_2, \cdots, w_n)$ can be determined. By the relation between distribution function and its characteristic function, f can be then uniquely determined. ■

Remark 2: In proposition 1, the purpose of the hypothesis that f is L^2-integrable is to assure the one-to-one relation between the density function and its characteristic function by Plancherel theorem [11]. The L^2-integrable hypothesis can involve a wealth of density functions, including almost all PDFs commonly used such as the exponential family and piecewisely smooth PDF with compact support.

Proposition 1 tells that any two PDFs with all the same one-dimensional marginal distributions are identical, which provides the theoretical basis for the possibility of learning PDF from its marginal distributions. The proof also implies a reconstruction scheme, in which the characteristic function's value at each point is determined firstly and then the inversion transform is carried out. Unfortunately, the scheme encounters two practical difficulties: one is that the characteristic function is hard to be computed in general and the other is that, usually, only a finite number of marginal distributions can be obtained. However, it is natural to expect that the distribution which can fit the given marginal distributions well in subspaces would be the actual distribution. For measuring the similarity of the marginal distribution of estimated PDF with the given marginal distribution, Kullback-Leibler distance, which is commonly used for evaluating the similarity of two density functions, is adopted. The definition of Kullback-Leibler distance is given as follows [12]:

Definition 3 (Kullback-Leibler distance): The Kullback-Leibler distance of density φ w.r.t. density ψ is defined by

$$D(\varphi||\psi) = \int \varphi \ln\left(\frac{\varphi}{\psi}\right). \quad (2)$$

[1]For higher dimensional subspaces or their mixtures, the proposition can still work in a similar way.

[2]Refer to [10] for the definition of the characteristic function.

Let X be the original space with the PDF estimation f, and X_k be a given subspace of X with a marginal distribution known as $g_{(k)}$ for $k = 1, \cdots, K$. Let $f_{(k)}$ denote the marginal distribution of f in X_k, and then the above problem can be formulated as the following optimization one:

$$f = \arg\min_{f \in \Omega} \sum_{k=1}^{K} D(g_{(k)}||f_{(k)}), \quad (3)$$

where Ω is the space of density functions f takes. Though the optimization reformulation seems reasonable, if there is no other restriction on Ω, it might fail to achieve the actual PDF because of serious ambiguity even all $D(g_{(k)}||f_{(k)})$ are minimized to zero (thus $f_{(k)} = g_{(k)}$). As the following proposition says, there can be an infinite number of density functions with the same marginal distributions in the finite subspaces arbitrarily given.

Proposition 4 (Nonuniqueness): Suppose f is a continuous density function in R^n $(n \geq 2)$, K is any natural number, the one dimensional subspace X_k $(k = 1, \cdots, K)$ is arbitrarily given and $f_{(k)}$ is the marginal distributions of f in X_k. There are infinitely density functions f' whose marginal distributions in X_k are $f_{(k)}$ for $k = 1, \cdots, K$.

Proof: The proof is accomplished by a construction method. Construct a function $g(\vec{x}) = g(x_1, \cdots, x_n)$ where $\vec{x} = (x_1, \cdots, x_n)^T \in R^n$ as follows:

$$g(\vec{x}) = \begin{cases} x_n(1 - \sum_{i=1}^{n} x_i^2), & \text{if } \sum_{i=1}^{n} x_i^2 < 1; \\ 0, & \text{otherwise.} \end{cases}$$

Obviously, g is a continuous function with compact support in R^n. The transform matrix from R^n into X_k is denoted by $P_k = (a_{1,(k)}, \cdots, a_{n,(k)})$, and P_k is extended to a full rank square matrix M_k with P_k the first row. Construct another function $\varphi_k(\vec{x}) = g(M_k\vec{x})$ for each k. Let

$$\phi(\vec{x}) = \varphi_1 * \varphi_2 * \cdots * \varphi_K$$

where $*$ represents the convolution operator. It is apparent that ϕ is still continuous with compact support. Besides, because f is continuous, then there exist a translation vector $\vec{t}_0$ and two positive scale factors λ_1, λ_2, *s.t.*

$$\forall \gamma \in [0, \lambda_2], f(\vec{x}) - \gamma\phi(\lambda_1(\vec{x} - \vec{t}_0)) \geq 0.$$

Let $f'_\gamma(\vec{x}) = f(\vec{x}) - \gamma\phi(\lambda_1(\vec{x} - \vec{t}_0))$, and it can be verified that f'_γ is still a density function and its marginal distribution in X_k is $f_{(k)}$. Since γ ranges from 0 to λ_2, the proof is completed. ■

Remark 5: Generally, it can be shown that proposition 2 still holds for f which is somewhere positive and continuous, which means that there is an open region where f is positive and continuous, e.g., uniformly distributions and piecewisely continuous PDFs.

However, if Ω only contains density functions of certain models with finite free parameters, such as Gaussian Mixture

Models and other finite mixture models of the exponential family [13], and if marginal distributions are properly given, the ambiguity will not occur. The verification depends on the detailed problem, but usually can be accomplished by solving equations. As most of the practical parametric models are of finite parameters, in many situations the actual PDF can be learned through the optimization task. Take GMM for example for further considerations. Additional notations are introduced first: Let the original space $X \subseteq R^d$, P_k be the transform matrix which projects X into X_k (X_k needs not to be one-dimensional and let its dimension be denoted by d_k), S be the sample set of X, and $\widehat{x}$ be an individual point in S. Then the projection of $\widehat{x}$ into X_k is $P_k\widehat{x}$, written by $\widehat{x}_{(k)}$. f is the underlying Gaussian mixtures:

$$f(x) = \sum_{l=1}^{c} \alpha_l h_l(x), \tag{4}$$

where $\sum_{l=1}^{c}\alpha_l = 1$, $\alpha_l \geq 0$, $h_l(x) \sim N(\mu_l, \Sigma_l)$[3] for $l = 1, 2, \cdots, c$, and $h_l(x)$ are distinct from each other. Therefore $f_{(k)} = \sum_{l=1}^{c}\alpha_l h_{l,(k)}(x)$ where

$$h_{l,(k)}(x) \sim N\left(P_k\mu_l, P_k\Sigma_l P_k^T\right). \tag{5}$$

Assume $\widehat{f}$ is another Gaussian mixture $\widehat{f}(x) = \sum_{\hat{l}=1}^{\hat{c}}\widehat{\alpha}_{\hat{l}}\widehat{h}_{\hat{l}}(x)$, where the symbols' meaning is defined in a similar way to f, and the marginal distributions in X_k are also $f_{(k)}$. Then $\widehat{h}_{\hat{l},(k)}(x)$ must be one of $h_{l,(k)}(x)$, i.e.,

$$(P_k\widehat{\mu}_{\hat{l}}, P_k\widehat{\Sigma}_{\hat{l}}P_k^T) \in \left\{\left(P_k\mu_l, P_k\Sigma_l P_k^T\right) | l = 1, \cdots, c\right\} \tag{6}$$

for $k = 1, \cdots, K$. If the transform matrices satisfy condition 1, by solving linear equations derived from (6), all feasible Gaussian components which can appear in $\widehat{f}$, including all $h_l(x)$, can be obtained finitely, written by $N(\xi_t, \Theta_t), t = 1, \cdots, T$, where T is the number of possible components obtained and the components are distinct with different t.

Condition 1: The dimensions of $span\{(P_k)_{i,\cdot}|k = 1, \cdots, K; i = 1, \cdots, d_k\}$[4] and $span\{$the upper triangle of $((P_k)_{i,\cdot})^T(P_k)_{j,\cdot}|i, j = 1, \cdots, d_k; k = 1, \cdots, K\}$ are d and $\frac{d(d+1)}{2}$, respectively.

Moreover, if the transform matrices satisfy condition 2, all the Gaussian components which are not in $\{h_l|l = 1, \cdots, c\}$ can be discarded also by (6).

Condition 2: (1) $\forall\xi \in \{\xi_t|t = 1, \cdots, T\}$, there is a transform matrix P s.t. $P\xi \notin \{P\xi_t|t = 1, \cdots, T; \xi_t \neq \xi\}$. (2) $\forall(\xi, \Theta) \in \{(\xi_t, \Theta_t)|t = 1, \cdots, T\}$, there exists a transform matrix Q such that $Q\Theta Q^T \notin \left\{Q\Theta_t Q^T|t = 1, \cdots, T; \xi_t = \xi \text{ but } \Theta_t \neq \Theta\right\}$.

[3]The notation as $h(x) \sim N(\mu, \Sigma)$ means that $h(x)$ is the Gaussion function: $h(x) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\}$, where d is the dimension of x.

[4]In the paper, for a matrix E, let $E_{i,\cdot}$, $E_{\cdot,j}$, E_{ij} stand for its i-th row vector, j-th column vector, entry in the i-th row and j-th column, respectively.

Till now, when condition 1 and 2 are both met, all the components in $\widehat{f}$ are identical to those of f and then all the coefficients. Though the combination of the above conditions is just one sufficient condition, they do indicate that if the subspaces are properly chosen, the original f can be achieved through the optimization task.

III. Algorithm for Learning GMM

In this section, the above objective function for learning GMM is optimized and the detailed optimization algorithm is developed. The meaning of the notations is the same as the previous section. In each subspace X_k, the marginal probability distribution, denoted by $g_{(k)}(x_{(k)})$, is obtained by prior knowledge or some estimation techniques. The problem is to obtain the Gaussian mixture f to minimize

$$\sum_{k=1}^{K} D(g_{(k)}||f_{(k)}). \tag{7}$$

In (7), though $D(g_{(k)}||f_{(k)})$ might be hard to compute directly, it can be approximated by the Monte Carlo integral. Samples with distribution $g_{(k)}$ can be generated by classical sampling methods such as MCMC [14], or directly use the projections of partial samples in $\widehat{X}$ chosen by bootstrap method [15] (thus $g_{(k)}$ needs not to be prior given or estimated). The later sampling method also helps to make full use of the samples with incomplete coordinates observed. Let the samples in X_k be denoted by $x_{i,(k)}$, $i = 1, \cdots, m_k$ (note that there is no relation between $x_{i,(k)}$ and $x_{i,(k')}$ for different k and k'). By Monte Carlo integration

$$D(g_{(k)}||f_{(k)}) \approx \frac{1}{m_k}\sum_{i=1}^{m_k}\log\frac{g_{(k)}(x_{i,(k)})}{f_{(k)}(x_{i,(k)})}, \tag{8}$$

eliminate the constant factors in (8) and obtain

$$D(g_{(k)}||f_{(k)}) \sim -\frac{1}{m_k}\sum_{i=1}^{m_k}\log f_{(k)}(x_{i,(k)}). \tag{9}$$

By substitution of (4)(9) in (7), the optimization objective is changed to maximize

$$\sum_{k=1}^{K}\left\{\frac{1}{m_k}\sum_{i=1}^{m_k}\log(\sum_{l=1}^{c}\alpha_l h_{l,(k)}(x_{i,(k)}))\right\}. \tag{10}$$

The objective function in (10) can not be optimized explicitly because it contains the log of the sum. We will utilize the idea of EM like that in [9], which has been proved to be able to provide a favorable convergence property for GMM [16]. However, in the following context, it can be found out that the EM algorithm can not be applied directly because the optima of the M-step is difficult to be obtained in closed form. Thus the generalized EM method (GEM, [1]), which does not require the global maxima in the M-step, is used instead. Let the latent random variable of $x_{i,(k)}$ be $z_{i,(k)} \in \{1, 2, \cdots, c\}$, whose value indicates the component in f "generating" $x_{i,(k)}$. Like the basic EM algorithm, GEM is an iterative algorithm

and the parameters should be initialized first. Let the estimation of $\alpha_l, \mu_l, \Sigma_l, f_l$ and f in the n-th iteration be denoted by $\alpha_l^{(n)}, \mu_l^{(n)}, \Sigma_l^{(n)}, f_l^{(n)}$ and $f^{(n)}$, respectively. Then

$$\begin{aligned} p_{i,(k)}^{(n)}(l) &\triangleq P(z_{i,(k)} = l | x_{i,(k)}, f^{(n)}) \\ &= \frac{\alpha_l^{(n)} f_{l,(k)}^{(n)}(x_{i,(k)})}{\sum_{j=1}^{c} \alpha_j^{(n)} f_{j,(k)}^{(n)}(x_{i,(k)})}. \end{aligned} \tag{11}$$

The expectation step (E-step):

$$\sum_{k=1}^{K} \frac{1}{m_k} \sum_{i=1}^{m_k} \sum_{l=1}^{c} p_{i,(k)}^{(n)}(l) \log(\alpha_l f_{l,(k)}(x_{i,(k)})). \tag{12}$$

Expand (12), and we have

$$\begin{aligned} \sum_{k=1}^{K} \sum_{l=1}^{c} \{ & r_{l,(k)}^{(n)} (\log \alpha_l - \frac{1}{2} \log |\Sigma_{l,(k)}| \\ & - \frac{1}{2} (P_k \mu_l)^T \Sigma_{l,(k)}^{-1} (P_k \mu_l)) + (P_k \mu_l)^T \Sigma_{l,(k)}^{-1} \overline{\theta}_{l,(k)}^{(n)} \\ & - \frac{1}{2} tr(\Sigma_{l,(k)}^{-1} \Gamma_{l,(k)}^{(n)}) \} + const \end{aligned} \tag{13}$$

where $r_{l,(k)}^{(n)} = \frac{1}{m_k} \sum_{i=1}^{m_k} p_{i,(k)}^{(n)}(l)$, $\Sigma_{l,(k)} = P_k \Sigma_l P_k^T$, $\overline{\theta}_{l,(k)}^{(n)} = \frac{1}{m_k} \sum_{i=1}^{m_k} p_{i,(k)}^{(n)}(l) x_{i,(k)}$, $\Gamma_{l,(k)}^{(n)} = \frac{1}{m_k} \sum_{i=1}^{m_k} p_{i,(k)}^{(n)}(l) x_{i,(k)} x_{i,(k)}^T$ and $tr(\cdot)$ is the matrix trace operator.

It is noticed that to maximize (13), each μ_l should satisfy

$$\mu_l = (\sum_{k=1}^{K} r_{l,k}^{(n)} P_k^T \Sigma_{l,(k)}^{-1} P_k)^{-1} (\sum_{k=1}^{K} P_k^T \Sigma_{l,(k)}^{-1} \overline{\theta}_{l,(k)}^{(n)}). \tag{14}$$

Moreover, Σ_l must be positive definite, so Σ_l can be decomposed into the form: $\Sigma_l = R_l^T \cdot R_l$, where R_l is a square matrix of $R_{d \times d}$. Let

$$\begin{aligned} H_1(R_l) &\triangleq \frac{1}{2} (\sum_{k=1}^{K} P_k^T (P_k R_l^T R_l P_k^T)^{-1} \overline{\theta}_{l,(k)}^{(n)})^T \\ &\quad \cdot (\sum_{k=1}^{K} r_{l,k}^{(n)} P_k^T (P_k R_l^T R_l P_k^T)^{-1} P_k)^{-1} \\ &\quad \cdot (\sum_{k=1}^{K} P_k^T (P_k R_l^T R_l P_k^T)^{-1} \overline{\theta}_{l,(k)}^{(n)}), \\ H_2(R_l) &\triangleq \frac{1}{2} \sum_{k=1}^{K} tr\{(P_k R_l^T R_l P_k^T)^{-1} \Gamma_{l,(k)}^{(n)}\}, \\ H_3(R_l) &\triangleq \frac{1}{2} \sum_{k=1}^{K} r_{l,(k)}^{(n)} \log |P_k R_l^T R_l P_k^T|, \end{aligned}$$

and then to maximize (12) is to maximize

$$\sum_{l=1}^{c} \{H_1(R_l) - H_2(R_l) - H_3(R_l) + \log \alpha_l \sum_{k=1}^{K} r_{l,(k)}^{(n)}\} \tag{15}$$

w.r.t. R_l, α_l for $l = 1, \cdots, c$. In the maximization step (M-step) for maximization of (15), the optima of α_l can be obtained by Lagrange multiplier method and given in (16), but the optima of R_l are difficult to obtained explicitly like the typical EM algorithm for GMM in [9]. Fortunately, the derivatives of $H_1(R_l), H_2(R_l), H_3(R_l)$ w.r.t. all entries in R_l can be derived (see appendix), and the derivatives of the objective value w.r.t. all entries in R_l can therefore be obtained, written by D_l, so the common gradient ascent method can be used. The incremental step size for each R_l, written by ε_l, is made variable and determined with the heuristic method which increases each ε_l individually from a small size by multiplying a fixed amplification factor till (15) stops increasing. The renewal formula are listed in (16) and the detailed algorithm are summarized in Table 1.

$$\begin{aligned} \alpha_l^{(n+1)} &= \frac{\sum_{k=1}^{K} r_{l,(k)}^{(n)}}{\sum_{l=1}^{c} \sum_{k=1}^{K} r_{l,(k)}^{(n)}}, \\ R_l^{(n+1)} &= R_l^{(n)} + \varepsilon_l D_l, \\ \Sigma_l^{(n+1)} &= (R_l^{(n+1)})^T \cdot R_l^{(n+1)}, \\ \mu_l^{(n+1)} &= (\sum_{k=1}^{K} r_{l,k}^{(n)} P_k^T (\Sigma_{l,(k)}^{(n+1)})^{-1} P_k)^{-1} \\ &\quad \cdot (\sum_{k=1}^{K} P_k^T (\Sigma_{l,(k)}^{(n+1)})^{-1} \overline{\theta}_{l,(k)}^{(n)}). \end{aligned} \tag{16}$$

TABLE I

THE DETAILED ALGORITHM FOR ESTIMATING GMM

■ Algorithm procedure

Input: Sample set $S_k = \{\hat{x}_{i,(k)} | i = 1, \cdots, m_k\}$ in X_k sampled from given marginal distributions $g_{(k)}$ for $k = 1, \cdots, K$.

Output: Parameter set $\Theta = \{(\alpha_l, \mu_l, \Sigma_l) \, | l = 1, \cdots, c\}$.

Step I : Initialize the number of components c, largest iteration steps N, iterating counter $n = 0$ and Θ. Decompose Σ_l: $\Sigma_l = R_l^T R_l$.

Step II: Iteration counter $n = n + 1$

–E-Step computes the $p_{i,(k)}^{(n)}(l)$ of each $\hat{x}_{i,(k)}$ by (11).

–D-Step computes the derivative D_l of (15) w.r.t. all entries in each R_l by the formulas in appendix.

–M-Step determines the step sizes ε_l and renews all R_l and Θ by (16).

Step III: Test whether the convergence is reached or $n > N$. If not, loop back to Step II, otherwise return Θ and exit.

IV. EXPERIMENTAL RESULTS

Experiments will be carried out on synthetic data and image segmentation in this section to verify the proposed algorithm, optimization framework and propositions. For comparison, the results of the EM for GMM will also be presented.

A. Synthetic Data

There are two datasets to be tested here: The first is uniformly sampled from the cirque,

$$0.7 \leq \sqrt{x_1^2 + x_2^2} \leq 1; \tag{17}$$

The second is generated from one complicated spiral manifold by

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} (2+10t)\sin(7\pi t) \\ (2+10t)\cos(7\pi t) \end{bmatrix} + 0.5 * \begin{bmatrix} n_1 \\ n_2 \end{bmatrix} \tag{18}$$

where t is uniformly distributed in $[0,1]$, and $n_1, n_2 \sim N(0,1)$. The component numbers of the algorithms in cirque and spiral datasets are respectively chosen to be 10 and 30 by experience. For the proposed algorithm, to meet condition 1 and 2, sufficient numbers of subspaces are set as 4 times the component numbers, 40 for the cirque and 120 for the spiral, and the transform matrices are randomly chosen on the unit circle with uniform distribution. The samples in marginal subspaces for the Monte Carlo integration are independently drawn from the original distributions and then projected by corresponding transform matrices, and all the sample numbers are set to be 1000. For the EM algorithm, the sample numbers are both set to be 20000. It is hard to measure performance quantitatively, so we provide the detailed results with the estimated component outlines in Fig. 1. It can be clearly observed that both algorithms can fit the manifolds well, however, our proposed approach only uses marginal distributions without any assumptions of the relations between different subspaces. Furthermore, the results show that it is possible to recover the original PDF merely from the distributions and the optimization task can achieve the underlying GMM. However, it should be noticed that, as a drawback, the computational cost of the proposed algorithm might be high due to the matrix inversions in the iterations.

B. Image Segmentation

The algorithms will be used for image segmentation. Two RGB colored images are tested, of which the sizes are 512×512 and 303×243, respectively. The original space adopted is the linear space of RGB color. The images will be segmented with 3 colors, so the numbers of components care all set to 3. For the proposed algorithm, the numbers of the subspaces $K = 40$ and all the subspaces are one-dimensional with the transform matrices randomly chosen from the unit sphere with uniform distribution. In each marginal subspace X_k, the samples are independently drawn from the images by bootstrap method and then projected by P_k, and all the sample numbers are set to be 1000. For the EM algorithm, all pixels in the images are used for estimation. Results are shown in Fig. 2. It is obvious that the results achieved by our method are satisfactory and as good as the EM algorithm.

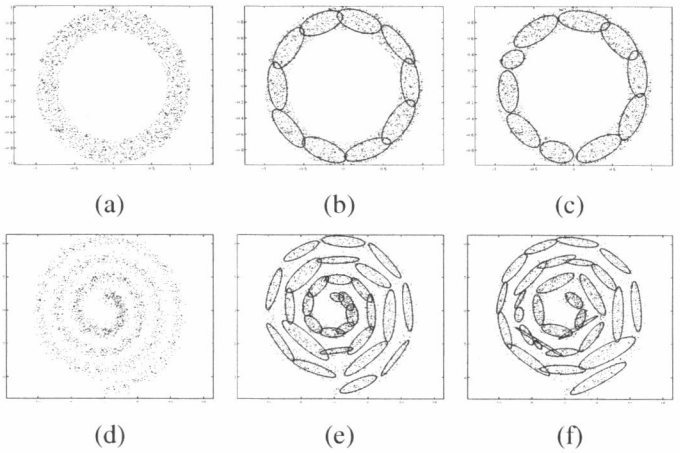

Fig. 1. Experimental results of synthetic data. The first column illustrates the original distributions, and the second column depicts the corresponding results by our algorithm, and the third column depicts the results by the EM algorithm for GMM.

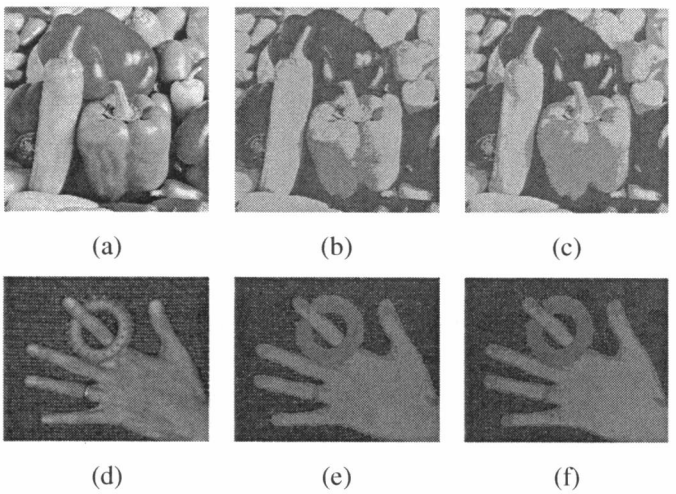

Fig. 2. Experimental results for image segmentation. The first column displays the original images, and the second column gives the results by our algorithm, and the third column depicts the results by the EM algorithm for GMM.

V. Conclusion and Future Work

This paper has shown that using marginal distributions to estimate the original probabilistic density function is theoretically possible for L^2-integrable density functions, and reformulated the recovery problem into an optimization task which is applicable for a wealth of density functions with finite parameters. For the application of the proposed formulation, it is employed on the important finite mixture models, Gaussian Mixture Models, and the corresponding optimization algorithm is developed. The proposed algorithm uses only the marginal distributions and imposes no assumptions on the relation between different subspaces. Compared with the EM algorithm based on whole coordinates of samples, experimental results show that the algorithm can perform well. The fact that the proposed algorithm can work well using only marginal

distributions in turn validates the proposed proposition and learning task. Furthermore, we can also expect that it would achieve promising performance when dealing with data which are partially visible in subspaces, since all these data can be used as samples in their visible subspaces for training without restrictions. To further the research, the problem of how to select subspaces properly for estimation and the weighted strategy for each KL distance in the objective function to boost the performance will be considered, and the optimization task will be employed on other mixture models and multivariate distributions.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their useful comments. This work is supported by the project (60475001) of the National Natural Science Foundation of China.

REFERENCES

[1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.
[2] D. W. Scott, *Multivariate Density Estimation*. New York: Wiley, 1992.
[3] J. Hwang, S. Lay, and A. Lippman, "Nonparametric multivariate density estimation: a comparative study," *IEEE Transactions on Signal Processing*, vol. 42, no. 10, pp. 2795–2810, 1994.
[4] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Statist.*, vol. 33, pp. 1065–1076, 1962.
[5] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, pp. 1–38, 1977.
[6] H. Greenspan, J. Goldberger, and A. Mayer, "Probabilistic space-time video modeling via piecewise GMM," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 26, no. 3, pp. 384–396, March 2004.
[7] R. J. Povinelli, M. T. Johnson, A. C. Lindgren, and J. Ye, "Time series classification using gaussian mixture models of reconstructed phase spaces," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 6, pp. 779–783, June 2004.
[8] S. Lucey and T. Chen, "A GMM parts based face representation for improved verification through relevance adaptation," in *2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04)*, vol. 2. Washington, D.C., USA: Springer, June 27–July 02 2004, pp. 855–861.
[9] J. A. Bilmes, "A gentle tutorial on the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," University of Berkeley, Tech. Rep. ICSI-TR-97-021, 1997.
[10] P. Billingsley, *Probability and Measure*, 3rd ed. New York: Wiley, 1995.
[11] W. Rudin, *Functional Analysis*, 2nd ed. New York: McGraw-Hill, 1991.
[12] T. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[13] G. J. McLachlan and D. Peel, *Finite Mixture Models*. New York: Wiley, 2000.
[14] C. Andrieu, N. D. Freitas, A. Doucet, and M. I. Jordan, "An introduction to MCMC for machine learning," *Machine Learning*, vol. 50, pp. 5–43, 2003.
[15] A. C. Davison and D. Hinkley, *Bootstrap Methods and Their Application*. Cambridge, New York: Cambridge University Press, 1999.
[16] L. Xu and M. I. Jordan, "On convergence properties of the EM algorithm for gaussian mixtures," *Neural Computation*, vol. 16, no. 1, pp. 129–151, January 1996.
[17] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*. Philadelphia: Society for Industrial and Applied Mathematics, 2000.
[18] B. Zhang, C. Zhang, and X. Yi, "Competitive EM algorithm for finite mixture models," *Pattern Recognition*, vol. 37, no. 1, pp. 131–144, 2004.

APPENDIX: THE MATRIX DERIVATIVES

The derivatives of $H_1(R_l), H_2(R_l), H_3(R_l)$ w.r.t. the entries in R_l can be achieved by the following steps. The basic knowledge on matrix derivative can be referred to [17]. Let $\Sigma_{l,(k)} \triangleq P_k R_l^T R_l P_k^T$, $V_l \triangleq \sum_{k=1}^{K} r_{l,(k)}^{(n)} P_k^T \Sigma_{l,(k)}^{-1} P_k$, $W_l \triangleq \sum_{k=1}^{K} P_k^T \Sigma_{l,(k)}^{-1} \overline{\theta}_{l,(k)}^{(n)}$, and F be an invertible matrix with F^* be its adjugate. Then $\frac{\partial F^{-1}}{\partial F_{ij}} = -\frac{1}{|F|^2}(F^*)_{ji}F^* + \frac{1}{|F|}\widetilde{F}$, where the entry $\widetilde{F}_{kl}$ in $\widetilde{F}$ is:

$$\widetilde{F}_{kl} = \begin{cases} 0, & \text{if } k = j \text{ or } l = i, \\ (-1)^{i+j}, & \text{if } F \in R_{2\times 2},\ k \neq j \text{ and } l \neq i, \\ (-1)^{i+j+k+l+I(l>i)+I(k>j)}|G^{(i,j,k,l)}|, & \text{otherwise} \end{cases}$$

where I is the indicator function and $G^{(i,j,k,l)}$ is obtained by reducing the i,l-th rows and j,k-th columns of F.

$$\begin{aligned}
\frac{\partial \Sigma_{l,(k)}^{-1}}{\partial (R_l)_{ij}} &= \sum_{s,t=1}^{d_m} \frac{\partial \Sigma_{l,(k)}^{-1}}{\partial (\Sigma_{l,(k)})_{st}} \{(R_l)_{i,\cdot}(P_k^T)_{\cdot,t}(P_k)_{sj} \\
&\quad + (P_k)_{s,\cdot}(R_l^T)_{\cdot,i}(P_k)_{tj}\}, \\
\frac{\partial V_l^{-1}}{\partial (R_l)_{ij}} &= \sum_{s,t=1}^{d}\sum_{k=1}^{K} \frac{\partial V_l^{-1}}{\partial (V_l)_{st}} r_{l,(k)}^{(n)} (P_k^T)_{s,\cdot} \frac{\partial \Sigma_{l,(k)}^{-1}}{\partial (R_l)_{ij}} (P_k)_{\cdot,t} \\
\frac{\partial W_l}{\partial (R_l)_{ij}} &= \sum_{k=1}^{K} P_k^T \frac{\partial \Sigma_{l,(k)}^{-1}}{\partial (R_l)_{ij}} \overline{\theta}_{l,(k)}^{(n)}, \\
\frac{\partial H_1(R_l)}{\partial (R_l)_{ij}} &= W_l^T V_l^{-1} \frac{\partial W_l}{\partial (R_l)_{ij}} + \frac{1}{2} W_l^T \frac{\partial V_l^{-1}}{\partial (R_l)_l} W_l, \\
\frac{\partial H_2(R_l)}{\partial (R_l)_{ij}} &= \frac{1}{2}\sum_{k=1}^{K} tr\{\frac{\partial \Sigma_{l,(k)}^{-1}}{\partial (R_l)_{ij}} \Gamma_{l,(k)}^{(n)}\}, \\
\frac{\partial H_3(R_l)}{\partial R_l} &= \frac{1}{2}\sum_{k=1}^{K} \frac{r_{l,(k)}^{(n)}}{|\Sigma_{l,(k)}|} R_l P_k^T \{\Sigma_{l,(k)}^* + (\Sigma_{l,(k)}^*)^T\} P_k.
\end{aligned}$$