Multimodal Data Representations with Parameterized Local Structures

Ying Zhu[1], Dorin Comaniciu[2],
Stuart Schwartz[1], and Visvanathan Ramesh[2]

[1] Department of Electrical Engineering, Princeton University
Princeton, NJ 08544, USA
{yingzhu, stuart}@EE.princeton.edu
[2] Imaging & Visualization Department, Siemens Corporate Research
755 College Road East, Princeton, NJ 08540, USA
{Dorin.Comaniciu, Visvanathan.Ramesh}@scr.siemens.com

Abstract. In many vision problems, the observed data lies in a nonlinear manifold in a high-dimensional space. This paper presents a generic modelling scheme to characterize the nonlinear structure of the manifold and to learn its multimodal distribution. Our approach represents the data as a linear combination of parameterized local components, where the statistics of the component parameterization describe the nonlinear structure of the manifold. The components are adaptively selected from the training data through a progressive density approximation procedure, which leads to the maximum likelihood estimate of the underlying density. We show results on both synthetic and real training sets, and demonstrate that the proposed scheme has the ability to reveal important structures of the data.

1 Introduction

In this paper we address the problem of learning the statistical representation of multivariate visual data through parametric modelling. In many pattern recognition and vision applications, an interesting pattern is measured or visualized through multivariate data such as time signals and images. Its random occurrences are described as scattered data points in a high-dimensional space. To better understand and use the critical information, it is important to explore the intrinsic low dimensionality of the scattered data and to characterize the data distribution through statistical modelling. The general procedure in learning parametric distribution models involves representing the data with a family of parameterized density functions and subsequently estimating the model parameters that best fit the data.

Among the commonly used parametric models, principal component analysis (PCA) [1, 2] and linear factor analysis [3] are linear modelling schemes that depict the data distribution either by a low-dimensional Gaussian or by a Gaussian with structured covariance. These approaches can properly characterize distributions in ellipsoidal shapes, but they are unable to handle situations where the

data samples spread into a nonlinear manifold that is no longer Gaussian. The nonlinear structure of a multivariate data distribution is not unusual even when its intrinsic variables are distributed unimodally. For instance, the images of an object under varying poses form a nonlinear manifold in the high-dimensional space. Even with a fixed view, the variation in the facial expressions can still generate a face manifold with nonlinear structures. A similar situation occurs when we model the images of cars with a variety of outside designs. The nonlinearity can be characterized by multimodal distributions through mixtured density functions. Such methods include local PCA analysis [5], composite analysis [4], transformed component and its mixture analysis [24, 25]. Alternative approaches have been proposed to describe the geometry of the principal manifold [6, 22, 23]. However, no probabilistic model is associated with the geometric representations.

This paper presents a new modelling scheme that characterizes nonlinear data distributions through probabilistic analysis. This scheme is built on parametric function representations and nonlinear factor analysis. Multivariate data is represented as a combination of parameterized basis functions with local supports. We statistically model the random parameterization of the local components to obtain a density estimate for the multivariate data. The probabilistic model derived here can provide likelihood measures, which are essential for many vision applications. We first introduced this idea in [26] to characterize the internally unimodal distributions with standard basis functions. Here we extend our discussion to cover multimodal distributions as well as the issue of basis selection. The paper is organized as follows. In section 2, we introduce the idea of parametric function representation. A family of multimodal distributions is formulated in section 3 to characterize an arbitrary data-generating process. In section 4, we solve the maximum likelihood (ML) density estimate through the procedure of progressive density approximation and the expectation-maximization (EM) algorithm. Related issues in basis selection and initial clustering are also addressed. In section 5, we show the experimental results of modelling both synthetic and real data. We finish with conclusions and discussions in section 6.

2 Data Representation with Parameterized Local Functions

Parameterized function representation [26] is built on function association and function parameterization (Fig. 1). In function association, an n-dimensional data $\mathbf{y} = [y_1, \cdots, y_n] \in R^n$ is associated with a function $y(\mathbf{t}) \in L^2 : R^d \rightarrow R^1 (\mathbf{t} \in R^d)$ such that

$$y_i = y(\mathbf{t}_i) \qquad (i = 1, \cdots, n)$$

For images, $\mathbf{t}$ is a two dimensional vector and $y(\mathbf{t})$ is defined on R^2 ($\mathbf{t} \in R^2$). For time signals, $\mathbf{t}$ becomes a scalar and $y(\mathbf{t})$ is defined on R^1 ($\mathbf{t} \in R^1$). The function $y(\mathbf{t})$ is interpolated from the discrete components of $\mathbf{y}$, and the vector $\mathbf{y}$ is produced by sampling $y(\mathbf{t})$. The function association applies a smoothing

process to the discrete data components. It is unique once the interpolation method is determined. Consequently, in the function space L^2, there exits a counterpart of the scattered multivariate data located in vector space R^n. The advantage of function association lies in its ease to handle the non-linearity by parametric effects embedded in the data. Its application to data analysis can be found in [6, 8–10, 19]. In function parameterization, we represent $y(\mathbf{t})$ with

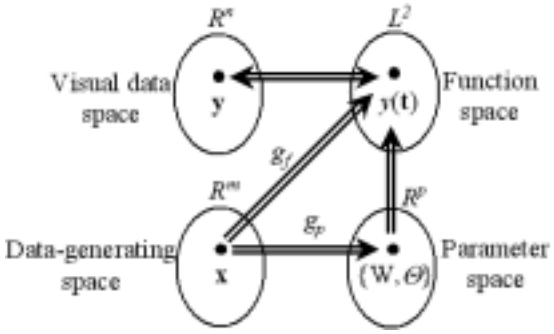


Fig. 1. Parametric function representation through space mappings. Multivariate data $\mathbf{y}$ is represented by function $y(\mathbf{t})$ parameterized by $\{W, \Theta\}$.

a basis of the function space L^2. Assume that the set of functions $\{b_\theta(\mathbf{t}) = b(\mathbf{t};\theta) : R^d \rightarrow R(\mathbf{t} \in R^d)\}$, each parameterized and indexed by θ, construct a basis of L^2. With the proper choice of $b(\mathbf{t};\theta)$, a function $y(\mathbf{t}) \in L^2$ can be closely approximated with a finite number (N) of basis functions,

$$y(\mathbf{t}) \cong [w_1, \cdots, w_N] \cdot \begin{bmatrix} b(\mathbf{t};\theta_1) \\ \vdots \\ b(\mathbf{t};\theta_N) \end{bmatrix} \tag{1}$$

In general, the basis function $b(\mathbf{t};\theta)$ is nonlinear in θ. If locally supported functions such as wavelets are chosen to construct the basis $\{b(\mathbf{t};\theta)\}$, then $y(\mathbf{t})$ is represented as a linear combination of nonlinearly parameterized local components. In the following discussion, we use W_N, Θ_N, and $\mathbf{\Theta}_N$ to denote the linear, nonlinear and overall parameter sets, where N is the number of basis functions involved.

$$W_N = \begin{bmatrix} w_1 \\ \vdots \\ w_N \end{bmatrix}; \Theta_N = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_N \end{bmatrix}; \mathbf{\Theta}_N = \begin{bmatrix} W_N \\ \Theta_N \end{bmatrix} \tag{2}$$

Assume $\mathbf{x} \in R^m$, in a vector form, to be the intrinsic quantities governing the data-generating process. With parametric function association, the observed data $\mathbf{y}$ is related to $\mathbf{x}$ through an unknown mapping $g_f : R^m \rightarrow L^2$, or equivalently, a mapping $g_p : R^m \rightarrow R^P$ from $\mathbf{x}$ to the parameter set,

$$g_p(\mathbf{x}) = [W_N(\mathbf{x}), \Theta_N(\mathbf{x})]^T \tag{3}$$

By defining the matrix

$$\mathbf{B}(\Theta_N) = [\mathbf{b}(\theta_1), \cdots, \mathbf{b}(\theta_N)]; \qquad \mathbf{b}(\theta_i) = \begin{bmatrix} b(\mathbf{t}_1; \theta_i) \\ \vdots \\ b(\mathbf{t}_n; \theta_i) \end{bmatrix} (i = 1, \cdots, N) \tag{4}$$

multivariate data $\mathbf{y}$ is related to $\mathbf{x}$ through the linear combination of local basis functions,

$$\begin{aligned} \mathbf{y} &= [y_1, \cdots, y_n]^T \\ &= [\mathbf{b}(\theta_1(\mathbf{x})), \cdots, \mathbf{b}(\theta_N(\mathbf{x}))] \cdot W_N(\mathbf{x}) + \mathbf{n}_N \\ &= \mathbf{B}(\Theta_N(\mathbf{x})) \cdot W_N(\mathbf{x}) + \mathbf{n}_N \end{aligned} \tag{5}$$

$\mathbf{n}_N$ is introduced to account for the noise in the observed data as well as the representation error in (1). By choosing a proper set of basis functions, (5) defines a compact representation for the multivariate data. Modelling the data distribution can be achieved through modelling the parameter set $\mathbf{\Theta}_N(\mathbf{x})$.

3 Learning Data Distribution

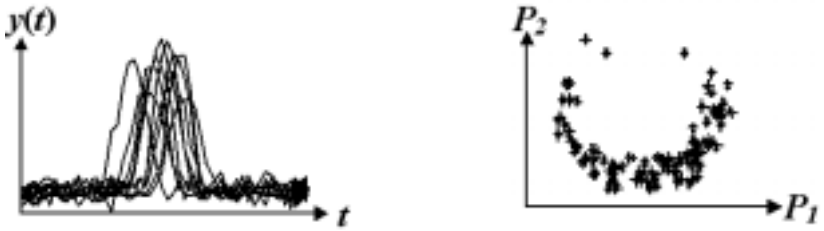
In this section, we discuss the algorithms and the criteria for learning nonlinear data distributions with parametric data representation. Fig. 2 shows an example of how the parametric effect can cause the nonlinear structure of the data distribution. The observed multivariate data $\mathbf{y} = [y_1, \cdots, y_n]^T$ consists of n equally spaced samples from the random realizations of a truncated raised cosine function $y_0(t)$,

$$\begin{aligned} y_i &= y(t_i; w, \theta) \qquad (t_i = i \cdot T) \\ y(t; w, \theta) &= w \cdot y_0(\tfrac{t-t_0}{s}) \qquad (\theta = (s, t_0)) \\ y_0(t) &= \begin{cases} \frac{1}{2}(1 + cos(t)) & (t \in [-\pi, \pi]) \\ 0 & (t \notin [-\pi, \pi]) \end{cases} \end{aligned} \tag{6}$$

where the generating parameters w, s and t_0 have a joint Gaussian distribution. Even though these intrinsic variables are distributed as a Gaussian, the conventional subspace Gaussian and Gaussian mixtures are either incapable or inefficient in describing the nonlinearly spread data. Such phenomena are familiar in many situations where the visual data is generated by a common pattern and bears similar features up to a degree of random deformation. Parameterized function representation decomposes the observed data into a group of local components with random parameters, which facilitates the characterization of locally deformed data.

3.1 Internally Unimodal Distribution

In most situations with a single pattern involved, the governing factor of the data-generating process is likely unimodal although the observed data $\mathbf{y}$ may



(a) (b)

Fig. 2. Nonlinearly distributed manifold. (a) Curve samples. (b) 2D visualization of the data distribution. (P_1 and P_2 signify the sample projections on the top two principal components derived from the data.)

disperse into a nonlinear manifold. For such an internally unimodal distribution, we assume a normal distribution for the intrinsic vector $\mathbf{x}$, which, together with a proper mapping g_p, generates $\mathbf{y}$. When the mapping g_p is smooth, the linearization of $W_N(\mathbf{x})$ and $\Theta_N(\mathbf{x})$ is valid around the mean of $\mathbf{x}$,

$$\begin{aligned} W_N(\mathbf{x}) &= W_{N,0} + A_{W,N} \cdot \mathbf{x} \\ \Theta_N(\mathbf{x}) &= \Theta_{N,0} + A_{\Theta,N} \cdot \mathbf{x} \end{aligned} \tag{7}$$

hence $W_N(\mathbf{x})$ and $\Theta_N(\mathbf{x})$ can be modelled as a multivariate Gaussian. Assume $\mathbf{n}_N$ is white Gaussian noise with zero mean and variance σ_N^2. From the representation (5), the multivariate data $\mathbf{y}$ is effectively modelled as

$$p(\mathbf{y}) = \int_{\mathbf{\Theta}_N} p(\mathbf{\Theta}_N) \cdot p(\mathbf{y}|\mathbf{\Theta}_N) d\mathbf{\Theta}_N \tag{8}$$

$$(\mathbf{y}|\mathbf{\Theta}_N = \mathbf{b}(\Theta_N) \cdot W_N + \mathbf{n}_N)$$

$$\mathbf{\Theta}_N \sim N(\mu_N, \Sigma_N); \qquad \mathbf{n}_N \sim N(\mathbf{0}, \sigma_N^2 \cdot I_n) \tag{9}$$

(8) defines a generative model of nonlinear factor analysis, where the parameter set $\mathbf{\Theta}_N(\mathbf{x})$ has a unimodal Gaussian distribution.

3.2 Multimodal Distribution

The adoption of multimodal distribution for $\mathbf{\Theta}_N(\mathbf{x})$ is necessary for two reasons. Firstly, if the process itself is internally multimodal, i.e. $\mathbf{x}$ can be modelled by mixtures of Gaussian, then the linearization of g_p around all the cluster means

$$\begin{aligned} W^q(\mathbf{x}) &= W_0^q + A_W^q \cdot \mathbf{x} \\ \Theta^q(\mathbf{x}) &= \Theta_0^q + A_\Theta^q \cdot \mathbf{x} \quad \mathbf{x} \in q\text{-th cluster} \qquad (q = 1, \cdots, C) \end{aligned} \tag{10}$$

leads to a mixture distribution for $\mathbf{\Theta}_N(\mathbf{x})$. Secondly, even in the case of an internally unimodal distribution, if the smoothness and the valid linearization of g_p do not hold over all the effective regions of $\mathbf{x}$, piecewise linearization of g_p is necessary, which again leads to a Gaussian mixture model for $\mathbf{\Theta}_N(\mathbf{x})$.

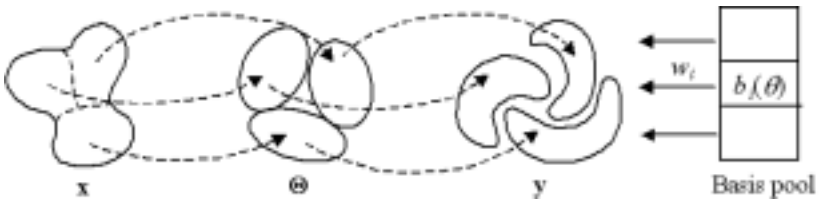


Fig. 3. Multimodal distribution and the basis pool for parameterized data representation.

Let c denote the cluster index, the generative distribution model for the observed data $\mathbf{y}$ is given by the multimodal factor analysis,

$$p(\mathbf{y}) = \sum_{q=1}^{C} P(c=q) \int_{\mathbf{\Theta}_{N_q}} p(\mathbf{\Theta}_{N_q}|c=q) \cdot p(\mathbf{y}|\mathbf{\Theta}_{N_q}, c=q) d\mathbf{\Theta}_{N_q} \tag{11}$$

$$P(c=q) = \pi_q \qquad (q = 1, \cdots, C) \tag{12}$$

$$p(\mathbf{\Theta}_{N_q}|c=q) = N(\mu_{q,N_q}, \Sigma_{q,N_q}) \tag{13}$$

$$p(\mathbf{y}|\mathbf{\Theta}_{N_q}, c=q) = N(\mathbf{B}(\Theta_{N_q}) \cdot W_{N_q}, \sigma^2_{q,N_q} \cdot I_n) \tag{14}$$

$P(c=q)$ denotes the prior probability for the q-th cluster, $p(\mathbf{\Theta}_{N_q}|c=q)$ denotes the density function of $\mathbf{\Theta}_{N_q}$ in the q-th cluster, and $p(\mathbf{y}|\mathbf{\Theta}_{N_q}, c=q)$ denotes the conditional density function of $\mathbf{y}$ given $\mathbf{\Theta}_{N_q}$ in the q-th cluster. Define $\Phi_{q,N_q} = \{\mu_{q,N_q}, \Sigma_{q,N_q}, \sigma^2_{q,N_q}\}$, $\mathbf{\Phi} = \{\pi_q, \Phi_{q,N_q}\}_{q=1}^{C}$. Equations (11)-(14) define a family of densities parameterized by $\mathbf{\Phi}$. The multivariate data $\mathbf{y}$ is statistically specified by the family of densities $p(\mathbf{y}|\mathbf{\Theta}, c)$. The parameters $\{\mathbf{\Theta}, c\}$, which characterize the cluster prior and the building components within each cluster, are specified by the family of densities $P(c)p(\mathbf{\Theta}|c)$ that depend on another level of parameters $\mathbf{\Phi}$. $\mathbf{\Phi}$ is therefore called the set of *hyper-parameters* [12]. The following discussion is devoted to finding the particular set of $\mathbf{\Phi}$ such that the generative distribution $p(\mathbf{y}|\mathbf{\Phi})$ best fits the observed data.

3.3 Learning through Maximum Likelihood (ML) Fitting

Given M independently and identically distributed data samples $\{\mathbf{y}_1, \cdots, \mathbf{y}_M\}$, the density estimate $\hat{p}(\mathbf{y})$, in the maximum likelihood (ML) sense, is then defined by the ML estimate of $\mathbf{\Phi}$ such that the likelihood

$$p(\mathbf{y}_1, \cdots, \mathbf{y}_M|\mathbf{\Phi}) = \prod_{i=1}^{M} p(\mathbf{y}_i|\mathbf{\Phi}) \tag{15}$$

is maximized over the parameterized family (11)-(14),

$$\begin{aligned} \hat{p}(\mathbf{y}) &= p(\mathbf{y}|\hat{\mathbf{\Phi}}) \\ \hat{\mathbf{\Phi}} &= \text{argmax}_{\mathbf{\Phi}\in\Omega} \prod_{i=1}^{M} p(\mathbf{y}_i|\mathbf{\Phi}) \end{aligned} \tag{16}$$

Ω denotes the domain of $\mathbf{\Phi}$. Further analysis suggests that the ML criterion minimizes the Kullback-Leibler divergence between the density estimate and the true density. Denote the true density function for the observed data by $p_T(\mathbf{y})$. The Kullback-Leibler divergence $D(p_T||\hat{p})$ measures the discrepancy between p_T and $\hat{p}$,

$$\begin{aligned} D(p_T||\hat{p}) &= \int p_T(\mathbf{y}) \cdot \log \frac{p_T(\mathbf{y})}{\hat{p}(\mathbf{y})} d\mathbf{y} \\ &= E_{p_T}[\log p_T(\mathbf{y})] - E_{p_T}[\log \hat{p}(\mathbf{y})] \end{aligned} \tag{17}$$

$D(p_T||\hat{p})$ is nonnegative and approaches zero only when the two densities coincide. Since the term $E_{p_T}[\log p_T(\mathbf{y})]$ is independent of the density estimate $\hat{p}(\mathbf{y})$, an equivalent similarity measurement is defined as

$$\begin{aligned} L(\hat{p}) &= E_{p_T}[\log \hat{p}(\mathbf{y})] \\ &= -D(p_T||\hat{p}) + E_{p_T}[\log p_T(\mathbf{y})] \\ &\leq E_{p_T}[\log p_T(\mathbf{y})] \end{aligned} \tag{18}$$

$L(\hat{p})$ increases as the estimated density $\hat{p}$ approaches the true density p_T. It is upper bounded by $E_{p_T}[\log p_T(\mathbf{y})]$. Since p_T is unknown, $L(\hat{p})$, the expectation of $\log \hat{p}(\mathbf{y})$, can be estimated in practice by its sample mean.

$$\hat{L}(\hat{p}) = \frac{1}{M} \sum_{i=1}^{M} \log \hat{p}(\mathbf{y}_i) \tag{19}$$

(19) defines the same target function as the ML estimate (16). Hence the ML fitting rule minimizes the Kullback-Leibler divergence between p_T and $\hat{p}$.

4 Hyper-parameter Estimation and Basis Selection

Fig. 4 illustrates the process of density estimation. To solve the ML density estimate, we need to construct the local basis functions $\{b_i\}$ as well as their parameter set $\mathbf{\Theta}$. For each cluster, the local building components are gradually introduced through the *progressive density approximation* procedure. With a set of local components $\{b_1, \cdots, b_N\}$ parameterized by $\mathbf{\Theta}_N$, the multivariate data $\mathbf{y}$ is represented as a linear combination of nonlinearly parameterized local components. The *expectation-maximization* (EM) algorithm [12] is applied to estimate the hyper-parameter set $\mathbf{\Phi}$ that defines the distribution of $\mathbf{\Theta}_N$, $\mathbf{n}_N$, and the cluster prior $P(c)$. We first discuss the EM algorithm, and then address the issues of basis selection and parameterization as well as initial clustering.

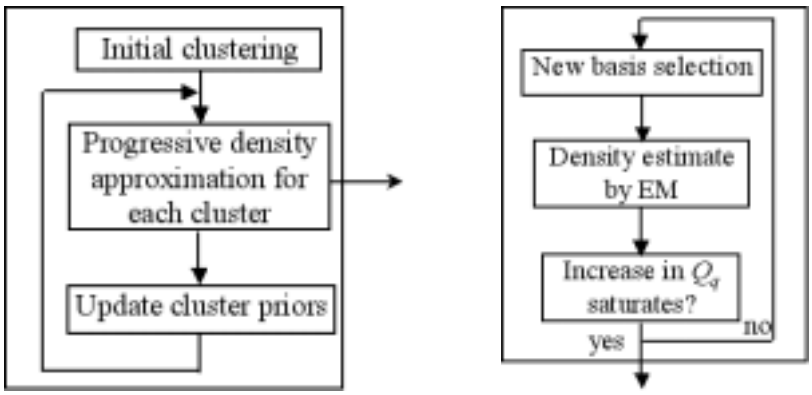


Fig. 4. Diagram of the EM-based modelling framework. (a) Iterative estimation of cluster prior and cluster density. (b) Progressive density approximation for each cluster.

4.1 EM Algorithm and Numerical Implementations

Assume that the set of local basis functions $\mathbf{B}_q(\Theta_{N_q}) = \{b_{q,1}(\theta_1), \cdots, b_{q,N_q}(\theta_{N_q})\}$ for the q-th cluster has already been established. Denote $\{\mathbf{\Theta}_{N,j}, c_j\}$ as the hidden parameter set for the observed data $\mathbf{y}_j$, and $\mathbf{\Phi}^{(k)} = \{\pi_q^{(k)}, \mu_{q,N_q}^{(k)}, \Sigma_{q,N_q}^{(k)}, \sigma_{q,N_q}^{2(k)}\}_{q=1}^{C}$ as the estimate of $\mathbf{\Phi}$ from the k-th step. The EM algorithm maximizes the likelihood $p(\mathbf{y}_1, \cdots, \mathbf{y}_M|\mathbf{\Theta})$ through the iterative expectation and maximization operations.

E-Step: Compute the expectation of the log-likelihood of the complete data $\log p(\{\mathbf{y}_j, \mathbf{\Theta}_{N,j}, c_j\}|\mathbf{\Phi})$ given the observed data $\{\mathbf{y}_j\}$ and the estimate $\mathbf{\Phi}^{(k)}$ from the last round,

$$Q(\mathbf{\Phi}|\mathbf{\Phi}^{(k)}) = \sum_{j=1}^{M} E[\log p(\mathbf{y}_j, \mathbf{\Theta}_{N,j}, c_j|\mathbf{\Phi})|\mathbf{y}_j, \mathbf{\Phi}^{(k)}] \tag{20}$$

M-Step: Maximize the expectation

$$\mathbf{\Phi}^{(k+1)} = \text{argmax}_{\mathbf{\Phi}} Q(\mathbf{\Phi}|\mathbf{\Phi}^{(k)}) \tag{21}$$

Denote

$$p_q(\mathbf{y}_j|\mathbf{\Theta}_{N_q,j}, \Phi_{q,N_q}) = p(\mathbf{y}_j|\mathbf{\Theta}_{N_q,j}, c_j = q, \Phi_{q,N_q}) \tag{22}$$

$$p_q(\mathbf{\Theta}_{N_q,j}|\Phi_{q,N_q}) = p(\mathbf{\Theta}_{N_q,j}|c_j = q, \Phi_{q,N_q}) \tag{23}$$

$$p_q(\mathbf{y}_j|\Phi_{q,N_q}) = p(\mathbf{y}_j|c_j = q, \Phi_{q,N_q}) \tag{24}$$

$$(q = 1, \cdots, C; \quad j = 1, \cdots, M)$$

$Q(\mathbf{\Phi}|\mathbf{\Phi}^{(k)})$ can be expressed as

$$Q(\mathbf{\Phi}|\mathbf{\Phi}^{(k)}) = \sum_{j=1}^{M}\sum_{q=1}^{C} P(c_j = q|\mathbf{y}_j, \mathbf{\Phi}^{(k)}) \log \pi_q + \sum_{q=1}^{C} Q_q(\Phi_{q,N_q}|\Phi_{q,N_q}^{(k)}) \quad (25)$$

where

$$Q_q(\Phi_{q,N_q}|\Phi_{q,N_q}^{(k)}) = \sum_{j=1}^{M} \frac{P(c_j = q|\mathbf{y}_j, \mathbf{\Phi}^{(k)})}{p_q(\mathbf{y}_j|\Phi_{q,N_q}^{(k)})} \int [\log p_q(\mathbf{y}_j|\mathbf{\Theta}_{N_q,j}, \Phi_{q,N_q}) + \log p_q(\mathbf{\Theta}_{N_q,j}|\Phi_{q,N_q})] \cdot p_q(\mathbf{y}_j|\mathbf{\Theta}_{N_q,j}, \Phi_{q,N_q}^{(k)}) p_q(\mathbf{\Theta}_{N_q,j}|\Phi_{q,N_q}^{(k)}) d\mathbf{\Theta}_{N_q,j} \quad (26)$$

(25) indicates that the cluster prior $\{\pi_q\}$ and the hyper-parameter set Φ_{q,N_q} for each cluster can be updated separately in the M-step. Generally, (26) has no closed form expression since the local component functions $\mathbf{B}_q(\Theta_{N_q})$ is nonlinear in Θ_{N_q}. The numerical approach can be adopted to assist the evaluation of the Q function. We detail the update rule for the hyper-parameters in the Appendix. The process of hyper-parameter estimation to maximize $p(\mathbf{y}_1, \cdots, \mathbf{y}_M|\mathbf{\Phi})$ is then summarized as follows:

1. Initially group $\{\mathbf{y}_1, \cdots, \mathbf{y}_M\}$ into C clusters. The initial clustering is addressed in later discussions. The number of clusters is preset. Set $\pi_q^{(0)} = \frac{M_q}{M}$, where M_q is the number of samples in the q-th cluster. Set $P(c_j = q|\mathbf{y}_j, \mathbf{\Phi}^{(0)})$ to 1 if $\mathbf{y}_j$ is assigned to the q-th cluster, 0 otherwise.
2. Construct local basis functions $\{b_1, \cdots, b_{N_q}\}$. Estimate the hyper-parameters $\{\Phi_{q,N_q}\}$ separately for each cluster. In later discussions, the progressive density approximation algorithm is proposed to gradually introduce local components and the EM algorithm is carried out to find the ML estimate of $\{\Phi_{q,N_q}\}$.
3. Use the EM procedure to iteratively update the cluster prior $\{\pi_q\}$ and the hyper-parameters $\{\Phi_{q,N_q}\}$ through (38)-(41) in the appendix .

4.2 Progressive Density Approximation and Basis Selection

Unlike other modelling techniques with a fixed representation, the proposed generative model actively learns the component functions to build the data representation. The procedure is carried separately for each cluster. By introducing more basis components, the density estimate gradually approaches the true distribution. The progressive density approximation for the q-th cluster is stated as follows:

1. Start with $Nq = 1$.
2. Find the ML density estimate $\hat{p}_q(\mathbf{y}) = p_q(\mathbf{y}|\hat{\Phi}_{q,N_q})$ by iteratively maximizing $Q_q(\Phi_{q,N_q}|\Phi_{q,N_q}^{(k)})$ with (39)-(41) in the appendix.
3. Introduce a new basis function, increase N_q by 1, and repeat step 2 and 3 until the increase of Q_q saturates as N_q increases.

Since the domain of Φ_{q,N_q} is contained in the domain of Φ_{q,N_q+1}, the introduction of the new basis function increases Q_q, $Q_q(\hat{\Phi}_{q,N_q+1}|\hat{\Phi}_{q,N_q}) \geq Q_q(\hat{\Phi}_{q,N_q}|\hat{\Phi}_{q,N_q})$, which leads to the increase of the likelihood $p(\mathbf{y}_1, \cdots, \mathbf{y}_M|\hat{\mathbf{\Phi}})$ and the decrease of the divergence $D(p_T||\hat{p})$.

Two issues are involved in basis selection. First, we need to choose the basis and its form of parameterization to construct an initial pool of basis functions (Fig. 3). Second, we need to select new basis functions from the pool for efficient data representation. Standard basis for the function space L^2, such as wavelets and splines with proper parameterization, is a natural choice to create the basis pool. In [26], we adopted wavelet basis (the Derivative-of-Gaussian and the Gabor wavelets) with its natural parameterization to represent the data:

Time signal $(y(t) : R \to R)$:

$$\begin{aligned} b(t;\theta) &= \psi_0(\tfrac{t-T}{s}); \\ \psi_0(t) &= t \cdot \exp(-\tfrac{1}{2}t^2);\ \theta = \{T, s\} \in R \times R^+ \end{aligned} \tag{27}$$

Image $(y(\mathbf{t}) : R^2 \to R)$:

$$\begin{aligned} b(\mathbf{t};\theta) &= \psi_0(SR_\alpha(\mathbf{t} - T) \\ \psi_0(\mathbf{t}) &= t_x \cdot \exp(-\tfrac{1}{2}(t_x^2 + t_y^2)) \quad (\mathbf{t} = [t_x, t_y]^T) \\ S = \begin{bmatrix} s_x & 0 \\ 0 & s_y \end{bmatrix}, &\quad R_\alpha = \begin{bmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{bmatrix}, \quad T = \begin{bmatrix} T_x \\ T_y \end{bmatrix} \\ \theta &= \{s_x, s_y, \alpha, T_x, T_y\} \in (R^+)^2 \times [0, 2\pi] \times R^2 \end{aligned} \tag{28}$$

The parameters naturally give the location, scale and orientation of the local components. Details on selecting new basis functions from the pool are provided in [26], where the new basis function is selected to maximize the Q function used by the EM procedure. Its actual implementation minimizes a term of overall residual energy evaluated with the current density estimate $p(\mathbf{y}|\hat{\Phi}_{q,N_q})$.

The standard basis provides a universal and overcomplete basis pool for all L^2 functions [14]. However, it does not necessarily give an efficient representation, especially when the data contains structures substantially different from the base function ψ_0. In this case, the representation can have a low statistical dimension but high complexity in terms of large number of basis functions involved. Here we propose an *adaptive basis selection* scheme that keeps the parameterization of the standard basis and replaces the base function ψ_0 by the *base templates* extracted from the data. Denote the n-th base template by $\bar{b}_{q,n}$, and use (28) to define the parameters. The building components are constructed as transformed base templates,

$$\begin{aligned} b_{q,n}(\mathbf{t};\theta) &= \bar{b}(Tr(\mathbf{t};\theta)) \\ Tr(\mathbf{t};\theta) &= SR_\alpha(\mathbf{t} - T) \end{aligned} \tag{29}$$

The hyper-parameter estimate $\hat{\Phi}_{q,N_q}$ with N_q components can be viewed as a special configuration of Φ_{q,N_q+1} where the (N_q+1)-th component is zero with probability 1. To select a new base template, $(w_{N_q+1}, \theta_{N_q+1})$ is initially assumed

to be independent of $\mathbf{\Theta}_{N_q}$ and uniformly distributed over its domain. From (37) and (22)-(24), we can approximate Q_q with the ML estimate of the sample parameters $(\hat{\mathbf{\Theta}}_{N_q,j}, \hat{w}_{N_q+1,j}, \hat{\theta}_{N_q+1,j})$,

$$Q_q(\Phi_{q,N_q+1}|\hat{\Phi}_{q,N_q}) \cong \kappa - \frac{1}{2\hat{\sigma}^2_{q,N_q}} \sum_{j=1}^{M} a_{q,j} \parallel \hat{r}_{N_q,j} - \hat{w}_{N_q+1,j} b_{q,N_q+1}(\hat{\theta}_{N_q+1,j}) \parallel^2$$

$$\hat{\mathbf{\Theta}}_{N_q,j} = \text{argmax}_{\mathbf{\Theta}_{N_q,j}} p_q(\mathbf{y}_j|\mathbf{\Theta}_{N_q,j}, \hat{\Phi}_{q,N_q}) p_q(\mathbf{\Theta}_{N_q,j}|\hat{\Phi}_{q,N_q})$$

$$\hat{r}_{N_q,j} = \mathbf{y}_j - \mathbf{B}(\hat{\Theta}_{N_q,j}) \cdot \hat{W}_{N_q,j}$$

$$a_{q,j} = \frac{P(c_j=q|\mathbf{y}_j, \mathbf{\Phi}^{(k)})}{p_q(\mathbf{y}_j|\Phi^{(k)}_{q,N_q})} p_q(\mathbf{y}_j|\hat{\mathbf{\Theta}}_{N_q,j}, \hat{\Phi}_{q,N_q}) p_q(\hat{\mathbf{\Theta}}_{N_q,j}|\hat{\Phi}_{q,N_q})$$

The new basis is selected to maximize Q_q, or equivalently, to minimize the weighted residue energy

$$\begin{aligned} &(\bar{b}_{q,N_q+1}, \{\hat{w}_{N_q+1,j}, \hat{\theta}_{N_q+1,j}\}_{j=1}^{M}) \\ &= \text{argmin} \sum_{j=1}^{M} a_{q,j} \parallel \hat{r}_{N_q,j} - \hat{w}_{N_q+1,j} b_{q,N_q+1}(\hat{\theta}_{N_q+1,j}) \parallel^2 \end{aligned} \tag{30}$$

From (29), the right hand term in (30) is evaluated as

$$\begin{aligned} &\sum_{j=1}^{M} a_{q,j} \parallel \hat{r}_{N_q,j}(\mathbf{t}) - \hat{w}_{N_q+1,j} \bar{b}_{q,N_q+1}(Tr(\mathbf{t}; \hat{\theta}_{N_q+1,j})) \parallel^2 \\ &= \sum_{j=1}^{M} a_{q,j} \hat{\omega}_j \parallel \hat{r}_{N_q,j}(Tr^{-1}(\mathbf{t}; \hat{\theta}_{N_q+1,j})) - \hat{w}_{N_q+1,j} \bar{b}_{q,N_q+1}(\mathbf{t}) \parallel^2 \\ &(Tr^{-1}(\mathbf{t}; \theta) = R_{-\alpha} S^{-1} \mathbf{t} + T; \quad \hat{\omega}_j = \hat{s}^{-1}_{x,q,N_q+1,j} \hat{s}^{-1}_{y,q,N_q+1,j}) \end{aligned} \tag{31}$$

The new basis selection procedure is stated as follows:

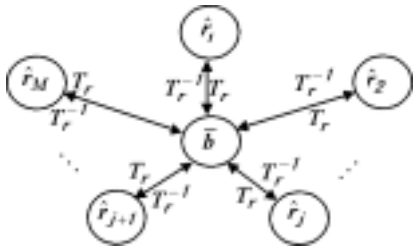


Fig. 5. Adaptive basis selection.

1. Locate the subregion where the residue $\sum_{j=1}^{M} a_{q,j} \parallel \hat{r}_{N_q,j}(\mathbf{t}) \parallel^2$ is most significant. Position the new base template to cover the subregion, and set $\bar{b}_{q,N_q+1}$ to be the subregion of $\hat{r}_{N_q,j_0}$ with $j_0 = \text{argmax}_j a_{q,j} \parallel \hat{r}_{N_q,j}(\mathbf{t}) \parallel^2$.

2. Iteratively update $\bar{b}_{q,N_q+1}$ and $(\hat{w}_{N_q+1,j}, \hat{\theta}_{N_q+1,j})$ to minimize (31).

$$(\hat{w}_{N_q+1,j}, \hat{\theta}_{N_q+1,j}) = \text{argmin}_{(w,\theta)} \parallel \hat{r}_{N_q,j}(\mathbf{t}) - w\bar{b}_{q,N_q+1}(Tr(\mathbf{t};\theta)) \parallel^2 \quad (32)$$

$$\bar{b}_{q,N_q+1}(\mathbf{t}) = \frac{1}{\sum_{j=1}^{M} a_{q,j}\hat{\omega}_j\hat{w}^2_{N_q+1,j}} \sum_{j=1}^{M} [a_{q,j}\hat{\omega}_j\hat{w}_{N_q+1,j}\hat{r}_{N_q,j}(Tr^{-1}(\mathbf{t};\hat{\theta}_{N_q+1,j}))] \quad (33)$$

3. Compute the hyper-parameters for Θ_{N_q+1} by (39)-(40), where $\mathbf{\Theta}^{(k)}_{N_q,j,i,q}$ is replaced by the ML estimate $\hat{\mathbf{\Theta}}_{N_q+1,j}$ and $K^{(k)}_{j,i,q}$ is replaced by the term derived from $\hat{\mathbf{\Theta}}_{N_q+1,j}$.

The base template and the ML sample parameters are derived simultaneously through iterative procedures (Fig. 5). The base template is updated by the weighted average of the inversely transformed samples, where samples with higher likelihood are weighted more. The sample parameters are updated by the transformation that most closely maps the base template to the sample.

4.3 Initial Clustering

Initial clustering groups together the data samples that share dominating global structures up to a certain transformation. Through the expansion

$$\bar{b}_{q,n}(Tr(\mathbf{t};\theta_{q,n})) = \bar{b}_{q,n}(\mathbf{t}) + [\bigtriangledown_{\mathbf{t}}\bar{b}_{q,n}(\mathbf{t})]^T \cdot (Tr(\mathbf{t};\theta_{q,n}) - \mathbf{t}) + \cdots \quad (34)$$

we notice that transformation effects are prominent in places where the local components have high-frequency (gradient) content. To emphasize the global structures for initial clustering, samples are smoothed by lowpass filters to reduce the effects of local deformations. Meanwhile, the sample intensity is normalized to reduce the variance of the linear parameters. Denote $\{\mathbf{y}_{s,1}, \cdots, \mathbf{y}_{s,M}\}$ as the smoothed and normalized data, the distance from $\mathbf{y}_{s,i}$ to $\mathbf{y}_{s,j}$ is defined as

$$d(\mathbf{y}_{s,i}, \mathbf{y}_{s,j}) = \min_{(w,\theta)\in\{(w_i,\theta_i)\}_i} \parallel \mathbf{y}_{s,j}(\mathbf{t}) - w_i\mathbf{y}_{s,i}(Tr(\mathbf{t};\theta_i)) \parallel^2 \quad (35)$$

where $\{(w_i, \theta_i)\}$ parameterize the global transformations. The cluster centers are first set to be the samples that have the minimum distance to a number of their nearest neighbors and that are distant from each other. Then the following procedure of clustering is performed iteratively until convergence.

1. Assign each sample to the cluster that has the minimal distance from its center to the sample.
2. For each cluster, find the new cluster center.

$$\mathbf{y}_{s,q} = \text{argmin}_{\mathbf{y}_{s,j}\in C_q} \sum_{\mathbf{y}_{s,i}\in C_q} d(\mathbf{y}_{s,j}, \mathbf{y}_{s,i}) \quad (36)$$

5 Experiments

Examples of progressive density approximation with standard basis have been shown in [26] to learn unimodal distributions, where wavelet basis is used to model curves as well as pose manifolds for object identification. Here we show two examples of learning multimodal distribution with adaptive basis selection.

5.1 Multimodal Distribution of Synthetic Data

In this experiment, 200 images with 30x30 pixels have been synthesized as the training data, each containing a cross with its vertical bar shifted randomly. The shift is around one of two centers shown by the examples in Fig. 6(a). The data cannot be efficiently modelled by the transformed component analysis [24] with only global transformations. The proposed modelling scheme has been performed to estimate the multimodal distribution with 2 clusters. Adaptive basis selection is implemented by positioning rectangular base templates to cover the regions with significant residue. The parameterization of the base template is defined by the horizontal translation and scaling parameters (s_x, T_x) (28). As shown in Fig. 6(b), the two cluster centers have been successfully recovered. In each cluster, three local components have been selected to represent the data. Meanwhile, the intrinsic dimension 1 for both clusters has been identified. Fig. 6(c) shows the synthesized realizations along the nonlinear principal manifold of each cluster.

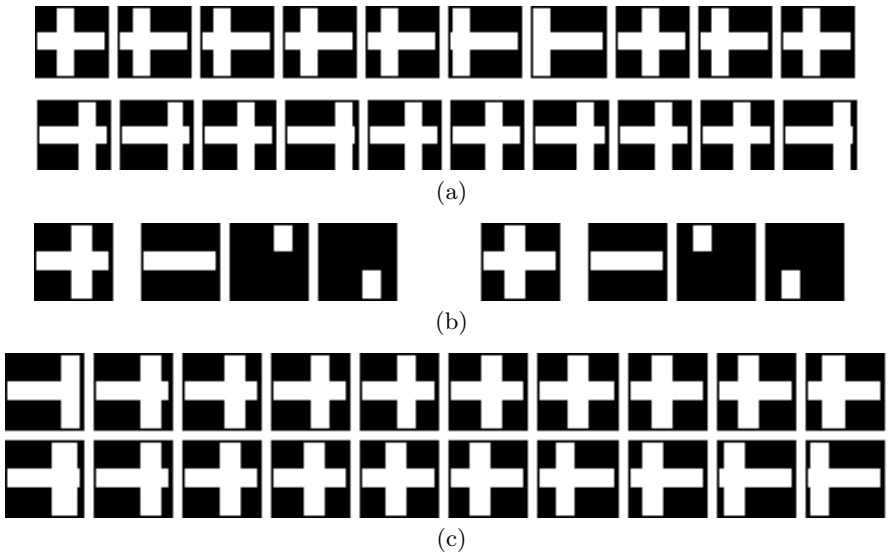


Fig. 6. Modelling synthetic data. (a) Training samples. (b) Cluster means and the three base templates selected for each cluster. (c) Synthesized samples on the principal manifolds learnt by the model. (one cluster in each row).

(a)

(b)

(c)

Fig. 7. Modelling mouths images. (a) Training samples defined by the subregions of the lower half face images. (b) From left to right: residue images, base templates and cluster means (one cluster in each row). (c) Images synthesized along the nonlinear principal components (the first and the second rows for the first cluster, the third and the fourth rows for the second cluster, and the last row for the third cluster).

5.2 Multimodal Distribution of Mouth Images

In this experiment, we are interested in modelling the mouth area with a multimodal distribution. A fixed region of 40x20 pixels was taken from the lower half of 100 face images to form the training set. Fig. 7(a) shows a few examples with open (smiling) and closed mouths. Since the region is fixed, there is no alignment information about the content inside it. We can extend the region for the entire face modelling. The parameterization defined in (28) is used with elliptical base templates adaptively selected to cover the areas with significant residue. The training images have been normalized before initial clustering. Three clusters and their nonlinear principal components identified by the model are shown in Fig. 7(b) and (c). The first cluster describes closed mouths from upright faces. Two local components have been selected to cover the mouth and the nose tip.

The second cluster describes closed mouths from slightly upward faces, and the third cluster describes the open mouths from smiling faces. Both the second and the third clusters use one component for the data representation. Fig. 7(c) indicates that horizontal and vertical translations are dominant deformations within the training set and they are represented by the nonlinear principal components.

6 Conclusions

In this paper, we have extended the idea of parameterized data representation to the statistical learning of multimodal data distributions. The building basis is adaptively selected from the training data to account for relevant local structures. The parameterized data representation by local components provides more flexibility than linear modelling techniques in describing the local deformations within the data. In addition, the EM-based generative model also provides a probabilistic description of the underlying data distribution. This allows various statistical approaches to be applied to vision problems. Both synthetic and real data are used to demonstrate the ability of the proposed modelling scheme to reveal the data structure and to obtain a good density estimate of the distribution manifold.

Through adaptive basis selection, the basis pool is adaptively defined by the data. It comprises the local patterns that are derived from the data. Compared with standard universal basis, the adaptive basis greatly reduces the complexity of the data representation. The algorithm finds, in a progressive and greedy fashion, the most efficient basis functions for the best modelling accuracy. Various applications of the proposed modelling scheme can be explored in further studies.

References

1. B. Moghaddam and A. Pentland, "Probabilistic Visual Learning for Object Representation", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, Jul. 1997, pp. 696-710.
2. M. Tipping and C. Bishop. "Probabilistic Principal Component Analysis". Technical Report NCRG/97/010, Neural Computing Research Group, Aston University, September 1997.
3. R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, Prentice Hall, NJ, 1992.
4. J. Ng and S. Gong, "Multi-view Face Detection and Pose Estimation Using A Composite Support Vector Machine Across the View Sphere", *RATFG-RTS*, 1999, pp. 14-21.
5. N. Kambhatla and T. K. Leen, "Dimension Reduction by Local PCA", *Neural Computation*, vol. 9, no. 7, Oct. 1997, pp. 1493-1516.
6. B. Chalmond and S. C. Girard, "Nonlinear Modeling of Scattered Multivariate Data and Its Application to Shape Change", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 21, no. 5, May 1999, pp. 422-434.

7. B. Moghaddam, "Principal Manifolds and Bayesian Subspaces for Visual Recognition", *IEEE Int. Conf. on Computer Vision*, 1999, pp. 1131-1136.
8. J. O. Ramsay and X. Li, "Curve Registration", *J. R. Statist. Soc.*, Series B, vol. 60, 1998, pp. 351-363.
9. G. James and T. Hastie, "Principal Component Models for Sparse Functional Data", Technical Report, Department of Statistics, Stanford University, 1999.
10. M. Black, and Y. Yacoob, "Tracking and Recognizing Rigid and Non-Rigid Facial Motions Using Local Parametric Models of Image Motion", *IEEE Int. Conf. Computer Vision*, 1995, pp. 374-381.
11. Z. R. Yang and M. Zwolinski, "Mutual Information Theory for Adaptive Mixture Models", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 4, Apr. 2001, pp. 396-403.
12. A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via EM Algorithm", *J. R. Statist. Soc.*, Series B, vol. 39, 1977, pp. 1-38.
13. H. Murase and S. K. Nayar, "Visual Learning and Recognition of 3D Objects from Appearance", *Int. J. Computer Vision*, vol. 14, 1995, pp. 5-24.
14. Q. Zhang and A. Benveniste, "Wavelet Networks", *IEEE Trans. Neural Networks*, vol. 3, no. 6, Nov 1992, pp. 889-898.
15. C. M. Bishop and J. M. Winn, "Non-linear Bayesian Image Modelling", *European Conf. on Computer Vision*, 2000, pp. 3-17.
16. B. Frey and N. Jojic, "Transformed Component Analysis: Joint Estimation of Spatial Transformations and image Components", *IEEE Int. Conf. Computer Vision*, 1999, pp. 1190-1196.
17. M. Weber, M. Welling and P. Perona, "Unsupervised Learning of Models for Recognition", *European Conf. on Computer Vision*, 2000, pp. 18-32.
18. T. S. Lee, "Image Representation Using 2D Gabor Wavelets", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 10, 1996, pp. 959-971.
19. B.W. Silverman, "Incorporating Parametric Effects into Functional Principal Components Analysis, *J. R. Statist. Soc.*, Series B, vol. 57, no. 4, 1995, pp. 673-689.
20. M. Black, and A. Jepson, "Eigentracking: Robust Matching and Tracking of Articulated Objects Using A View-based Representation", *European Conf. on Computer Vision*, 1996, pp. 329-342.
21. A. R. Gallant, *Nonlinear Statistical Models*, John Wiley & Sons Inc., NY, 1987.
22. J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction", *Science*, vol. 290, 2000, pp. 2319-2323.
23. S. Roweis and L. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding", *Science*, vol. 290, 2000, pp. 2323-2326.
24. B. J. Frey and N. Jojic, "Transformation-Invariant Clustering and Dimensionality Reduction Using EM", submitted to *IEEE Trans. Pattern Analysis and Machine Intelligence*, Nov. 2000.
25. C. Scott and R. Nowak, "Template Learning from Atomic Representations: A Wavelet-based Approach to Pattern Analysis", *IEEE workshop on Statistical and Computational Theories of Vision*, Vancouver, CA, July 2001.
26. Y. Zhu, D. Comaniciu, Visvanathan Ramesh and Stuart Schwartz, "Parametric Representations for Nonlinear Modeling of Visual Data", *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2001, pp. 553-560.
27. K. Popat and R. W. Picard, "Cluster Based Probability Model and Its Application to Image and Texture Processing", *IEEE Trans. Image Processing*, Vol. 6, No. 2, 1997, pp. 268-284.

Appendix

Using the idea of importance sampling, for each cluster, a group of random realizations of $\mathbf{\Theta}_{N_q,j}$, $\mathbf{\Theta}^{(k)}_{N_q,j,i,q} = [W^{(k)}_{N_q,j,i,q}, \Theta^{(k)}_{N_q,j,i,q}]^T_i$, is chosen within the volume where the value of $p_q(\mathbf{y}_j|\mathbf{\Theta}_{N_q,j}, \Phi^{(k)}_{q,N_q}) p_q(\mathbf{\Theta}_{N_q,j}|\Phi^{(k)}_{q,N_q})$ is significant, and Q_q is evaluated as

$$Q_q(\Phi_{q,N_q}|\Phi^{(k)}_{q,N_q}) \cong \sum_{j=1}^{M} K^{(k)}_{j,i,q}[\log p_q(\mathbf{y}_j|\mathbf{\Theta}^{(k)}_{N_q,j,i,q}, \Phi_{q,N_q}) + \log p_q(\mathbf{\Theta}^{(k)}_{N_q,j,i,q}|\Phi_{q,N_q})]$$

$$K^{(k)}_{j,i,q} = \frac{P(c=q|\mathbf{y}_j, \Phi^{(k)}_{q,N_q})}{p_q(\mathbf{y}_j|\Phi^{(k)}_{q,N_q})} p_q(\mathbf{y}_j|\mathbf{\Theta}^{(k)}_{N_q,j,i,q}, \Phi^{(k)}_{q,N_q}) p_q(\mathbf{\Theta}^{(k)}_{N_q,j,i,q}|\Phi^{(k)}_{q,N_q}) \cdot \kappa^{(k)}_{q,j}$$

$$\kappa^{(k)}_{q,j} = \frac{p_q(\mathbf{y}_j|\Phi^{(k)}_{q,N_q})}{\sum_i p_q(\mathbf{y}_j|\mathbf{\Theta}^{(k)}_{N_q,j,i,q}, \Phi^{(k)}_{q,N_q}) p_q(\mathbf{\Theta}^{(k)}_{N_q,j,i,q}|\Phi^{(k)}_{q,N_q})} \tag{37}$$

Substitute the density function in (37) with (12)-(14), the cluster prior $\{\pi_q\}$ and the hyper-parameter set Φ_q for each cluster are updated separately in the M-step.

$$\pi^{(k+1)} = \frac{\sum_{j=1}^{M} P(c=q|\mathbf{y}_j, \Phi^{(k)}_{q,N_q})}{\sum_{q=1}^{C}\sum_{j=1}^{M} P(c=q|\mathbf{y}_j, \Phi^{(k)}_{q,N_q})} \tag{38}$$

$$\mu^{(k+1)}_{q,N_q} = \frac{1}{\sum_{j=1}^{M}\sum_i K^{(k)}_{j,i,q}} \sum_{j=1}^{M}\sum_i K^{(k)}_{j,i,q} \mathbf{\Theta}^{(k)}_{N_q,j,i,q} \tag{39}$$

$$\Sigma^{(k+1)}_{q,N_q} = \frac{1}{\sum_{j=1}^{M}\sum_i K^{(k)}_{j,i,q}} \sum_{j=1}^{M}\sum_i K^{(k)}_{j,i,q} \cdot (\mathbf{\Theta}^{(k)}_{N_q,j,i,q} - \mu^{(k+1)}_{q,N_q}) \cdot (\mathbf{\Theta}^{(k)}_{N_q,j,i,q} - \mu^{(k+1)}_{q,N_q})^T \tag{40}$$

$$\sigma^{2(k+1)}_{q,N_q} = \frac{1}{Card(\mathbf{y}) \cdot \sum_{j=1}^{M}\sum_i K^{(k)}_{j,i,q}} \sum_{j=1}^{M}\sum_i K^{(k)}_{j,i,q} \parallel \mathbf{y}_j - \mathbf{B}_q(\Theta^{(k)}_{N_q,j,i,q}) \cdot W^{(k)}_{N_q,j,i,q} \parallel^2 \tag{41}$$

$Card(\mathbf{y})$ denotes the cardinality of the multivariate data $\mathbf{y}$.