Mixtures of Principal Components Gaussians for Density Estimation in High Dimension Data Spaces

Lemarié Bernard

La Poste SRTP 10, Rue de l'Île-Mabon BP 86334 44263 Nantes Cedex 02 FRANCE
E-mail : lemarie@srtp.srt-poste.fr

Abstract : An approximation of the gaussian model for density estimation in high-dimension data spaces is presented. The work is mainly motived by the need for a numerically tractable model in high dimension data spaces. The characteristic of the model is to restrict to the Principal Component of each local covariance matrix with the advantage of still being a probabilistic model. The likelihood of the local density is studied and an iterative algorithm is next proposed so as to learn the model. This latter is an adaptation of the well-known iterative Generalized Hebbian Algorithm. A comparison is done with related works based on Factorial Analysis. First experiments on handwritten digits are also reported.

Keywords: Density estimation, Approximation by mixtures, Handwritten recognition.

1 Introduction.

Mixture of gaussians is a well-known and widely used method for density estimation, partly thanks to its Expectation Maximization (EM) based learning algorithm. However, when dealing with high dimension data spaces some difficulties appear due to the sparseness of the data. In that case the likelihood of a pattern mostly reduces to the nearest local mixture along with an exponential decrease. Also, the important number of parameters may require high computation time as well a large database for learning. Finally, such systems present poor generalization ability and some attempts have been proposed in order to improve it for example by adding a term of regularization [12]. On the other hand, a convenient method to reduce space dimension is the Principal Component Analysis (PCA). For example in the Karhunen-Loeve Transform (KLT), data are projected onto the principal eigenvectors subspace and it is a well known result that this projection is optimal in term of the distance between a pattern and its projection, also called reconstruction error.The earlier methods for PCA were generally used in a batch context and more recently, especially in the field of neural networks, various iterative methods have been proposed [9], [11], [13]. From a practical viewpoint these algorithms are considered as efficient and useful in the case of a data evolving context.

We have presented recently the basic ideas for a combination of Mixture and PCA [8]. We called the model M-PCG which stands for Mixtures of Principal Components Gaussians. In fact, if the problem of dimensionality reduction in the framework of mixture models have yet been addressed by several authors [4],[6], we have not found, except a parallel work [15] with a Factorial Analysis (FA) approach, any example of the association of the mixture model and PCA in a fully statistical context. Our objective is this paper is thus to detail this model and its recent evolutions, to compare it with parallel works on FA, and also to illustrate its interest with an example of classifying handwritten digits.

The main feature of the presented model is that in spite of the restriction to a few components i.e. to a few directions in the data space, it remains a statistical model expressed by a true probabilistic local density which can be included in the general statistical modelling of the problem to solve. Basically, thanks to this probabilistic expression, the local restriction to the principal components remains directly comparable to others local components restriction in different directions.This comparison is facilitated by the use of a global parameter which also offers a mean to regularize the mixture.

More formally, we will show, that under some conditions this local probability density is justified in term of Maximum Likelihood Estimation (MLE). Once this proof established, a classical and convergent training method for the M-PCG model could be based on the EM algorithm [1],[16] with at each step a complete extraction of the principal eigenvectors. In practice, this solution is not appealing because the computing of the principal components at each step might be time consuming and not adapted to a data evolving environment or incremental use of the model in the sense proposed in [10]. We thus propose, without proof of convergence, to use an iterative method for principal components extraction.

Combined with attractive properties in term of computation time, the model thus presents many applications such as data compression, density estimation and statistical classification. Our field of interest is the handwritten recognition with handwritten character classification but also the Continuous Density Hidden Markov Model for handwritten words recognition [7]. Also, several recent

works in the field of handwritten character modelling are based on autoencoder networks [4] and we will connect these works to our approach.

The paper is organized as follows. The second chapter presents the statistical model. The chapter 3 studies the optimality of the model and the chapter 4 considers the training of the model. The section 5 establishes links of our model with recent works.The section 6 reports the application of the model to a task of character classification. The final chapter presents the conclusion and some perspectives.

2 The M-PCG Model.

Given a point $x \in R^d$, the mixture is a weighted sum of local densities:

$$p(x|\Theta) = \sum_{j=1}^{M} \alpha_j g(x|\Theta_j) \qquad with \qquad \sum_{j=1}^{M} \alpha_j = 1$$

In the M-PCG model, the local density $g(x|\Theta_j)$ corresponds to a classical multi-dimensional gaussian with the restriction of the lowest eigenvalues to a default width parameter σ^2. By decomposing the local covariance matrix among its eigenvectors this local probability can be expressed by:

$$g(x|\Theta_j) = \frac{\exp\left(-\frac{1}{2}\sum_{k=1}^{P}\frac{((x-u_j)^T V_{jk})^2}{\lambda_{jk}}\right)}{\prod_{k=1}^{P}\sqrt{2\Pi}\sqrt{\lambda_{jk}}} \times \frac{\exp\left(-\frac{1}{2}\sum_{k=P+1}^{d}\frac{((x-u_j)^T V_{jk})^2}{\sigma^2}\right)}{\prod_{k=P+1}^{d}\sqrt{2\Pi}\sigma}$$

where u_j is the centre of the local density. The sets $\{\lambda_{j1}, ..., \lambda_{jd}\}$ and $\{V_{j1}, ..., V_{jd}\}$ are the sets of respectively eigenvalues and eigenvectors of the local covariance matrix Σ_j, ordered by $\lambda_{jv} \leq \lambda_{ju}$ if $1 \leq u \leq v \leq d$. P is the selected number of principal eigenvalues.The model remains thus very close to the classical multi-dimensional gaussian model. In particular the M-PCG model still works in all the directions of the data space and the function $g(x|\Theta_j)$ is well stochastic.

If we note x_j^p the projection of the pattern onto the local P-dimensional principal eigen-subspace we interpret he first term as a local deformation term which takes into account the weighted distance between the local centre and this point. We can also write:

$$\sum_{k=P+1}^{a} ((x-u_j)^T V_{jk})^2 = d(x, x_j^p)^2$$

and then the second term is an orthogonal term which depends on the distance between the pattern and its projection x_j^p. We can remark that the distance in the orthogonal term is nothing but the cost of reconstruction of the Karhunen-Loeve Transform. The essential difference between this latter and our model is thus the use of the deformation term and the inclusion of the cost in a probabilistic model. An analogue remark can be done if we compare our model with the Constrained Rank Gaussian Mixture Bayes classifier introduced in [6] which this time uses only the deformation term but not the orthogonal term.

A last formulation can also be considered with:

$$d(x, x^p)^2 = d(x, u_j)^2 - d(x^p, u_j)^2 = d(x, u_j)^2 - \sum_{k=1}^{P} ((x-u_j)^T V_{jk})^2$$

which gives:

$$g(x|\Theta_j) = \frac{\exp\left(-\frac{1}{2}\frac{d(x,u_j)^2}{\sigma^2}\right)}{(\sqrt{2\Pi})^d \sigma^{d-P}} \times \frac{\exp\left(-\frac{1}{2}\sum_{k=1}^{P}\left(\frac{1}{\lambda_{jk}} - \frac{1}{\sigma^2}\right)((x-u_j)^T V_{jk})^2\right)}{\prod_{k=1}^{P}\sqrt{\lambda_{jk}}}$$

This expression of the local density shows that we do not need to know all the eigenvalues for computing the local density. Only the P scalar products and the distance to the local centre are required.

In this work, we have chosen an unique width parameter σ for the orthogonal density as well as an unique parameter P for the number of principal components estimation. With this rule, the normalization term $1/(\sqrt{2\Pi}\sigma)^{d-P}$ can be factorized among all the units, avoiding thus the reduction of the estimation to the most likely centre i.e. to a Winner Take All scheme. We do believe that this is

an important aspect of high dimensional density estimation especially when separately built mixtures must be compared. Finally, since the number of components as well as the number of mixtures are not learned during the training process, we have for the model the following parameters:

$$\Theta = \bigcup_{j=1}^{M} \{u_j, \alpha_j, V_{j1}, ..., V_{jP}, \lambda_{j1}, ..., \lambda_{jP}\} \cup \{\sigma\}$$

3 Maximum Likelihood Estimation.

In this section, we show that the local density model PCG is optimal in the sense of the likelihood estimation. Note that, as a global parameter, σ is fixed in this step. The demonstration is based on classical results of eigenvalues analysis but to our knowledge have not yet been published. Given a point $w \in R^d$ and an orthonormal basis $\{W_1, ..., W_P\}$ of vectors of a P-dimensional subspace S and given a set $\{\mu_1, ..., \mu_P\}$ of positive numbers, we write the local density as:

$$g(x|\Theta) = \frac{\exp\left(-\frac{1}{2}\frac{d(x,w)^2}{\sigma^2}\right)}{(\sqrt{2\Pi}\sigma)^{d-P}} \times \frac{\exp\left(-\frac{1}{2}\sum_{k=1}^{P}\left(\frac{1}{\mu_k}-\frac{1}{\sigma^2}\right)((x-w)^T W_k)^2\right)}{\prod_{k=1}^{P}\sqrt{2\Pi}\sqrt{\mu_k}}$$

Given a data set $X = \{x_1, x_2, ..., x_N\} \subset R^d$, the likelihood of the model is:

$$L(\Theta|X) = \sum_{i=1} \ln(g(x_i|\Theta)) = -\frac{N}{2}((d-P)\ln(\sigma^2) + \ln(2\Pi))$$

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{N} d(x_i, w)^2 - \frac{N}{2}\sum_{k=1}^{P}\ln(\mu_k) - \frac{1}{2}\sum_{k=1}^{P}\left(\frac{1}{\mu_k}-\frac{1}{\sigma^2}\right)\sum_{i=1}^{N}((x_i-w)^T W_k)^2$$

We have to show that the likelihood is maximized when the sets $\{W_1, ..., W_P\}$ and $\{\mu_1, ..., \mu_P\}$ correspond to the sets $\{V_1, ..., V_P\}$ and $\{\lambda_1, ..., \lambda_P\}$ of the principal eigenvectors and associated eigenvalues. The derivation with respect to the centre of the mixture is nearly the same than in the case of the multidimensional gaussians mixture and leads to a maximum at the mass centre u of the data. On the other hand, since the euclidean distance is independent of the orthonormal basis used, the dependence of the likelihood to the vectors $\{W_1, ..., W_P\}$ is only through the last term of the above formula. Moreover we have:

$$\sum_{i=1}^{N}((x_i-u)^T W_k)^2 = W_k{}^T\left(\sum_{i=1}^{N}(x_i-u)(x_i-u)^T\right)W_k = N \times W_k{}^T \Sigma W_k$$

where Σ is the covariance matrix. The analysis of the optimality is thus very close to the analysis for the case of the reconstruction error in the KLT expansion with the same orthonormality constraints. For example in [2], these constraints are first introduced in the term to maximize through Lagrange multipliers so as to produce the necessary condition for the subspace S to be invariant by the covariance matrix. Next, it is classically shown by considering the restricted local covariance matrix that this subspace is an eigen-subspace, that is a subspace generated by a subset $U_{l_1}, ..., U_{l_P}$ $1 \le l_k \le P \quad \forall k = 1, ..., P$ of the eigenvectors of the covariance matrix. This demonstration immediately transposes to our case under the hypothesis $\mu_k \neq \sigma^2$, $\forall k = 1, ..., P$. We can now write:

$$w_{kl_u} = w_{ul_k} \qquad \forall k = 1, ..., P \qquad \forall u = 1, ..., P$$

$$W_k = \sum_{u=1}^{P} w_{kl_u} U_{l_u} \qquad \sum_{k=1}^{P} w_{kl_u}^2 = \sum_{k=1}^{P} w_{ul_k}^2 = 1 \qquad \forall u = 1, ..., P$$

The next necessary condition concerns the parameter $\mu_k \qquad \forall k = 1, ..., P$:

$$\frac{\partial}{\partial \mu_k} L(\Theta|X) = 0 \Rightarrow \mu_k = W_k{}^T \Sigma W_k = \sum_{u=1}^{P} \lambda_{l_u} w_{kl_u}^2$$

Thus if W_k is an eigenvector, μ_k is well the associated eigenvalue. For a given S, the maximization of the likelihood is now equivalent to the maximization of:

$$-\frac{N}{2}\sum_{k=1}^{P}\ln(W_k^T \Sigma W_k) - \frac{N}{2}\sum_{k=1}^{P}\left(\frac{1}{W_k^T \Sigma W_k} - \frac{1}{\sigma^2}\right) W_k^T \Sigma W_k = -\frac{N}{2} - \frac{N}{2}\sum_{k=1}^{P}\ln\left(\sum_{u=1}^{P}\lambda_{l_u} w_{kl_u}^2\right) + \frac{N}{2\sigma^2}\sum_{k=1}^{P}\sum_{u=1}^{P}\lambda_{l_u} w_{kl_u}^2$$

The third term is the trace of the local restricted covariance matrix which is easily shown to be invariant to the choice of the orthonormal basis $\{W_1, ..., W_P\}$ of S. For the second term, the convexity of the logarithm function gives:

$$-\sum_{k=1}^{P} \ln\left(\sum_{u=1}^{P} \lambda_{l_u} w_{kl_u}^2\right) < -\sum_{k=1}^{P}\sum_{u=1}^{P} w_{kl_u}^2 \ln(\lambda_{l_u}) = -\sum_{u=1}^{P} \ln(\lambda_{l_u})\left(\sum_{k=1}^{P} w_{kl_u}^2\right) = -\sum_{u=1}^{P} \ln(\lambda_{l_u})$$

This last value being precisely obtained when the set $\{W_1, ..., W_P\}$ is the set of the eigenvectors. Therefore, for a given eigen-subspace S the optimal value is reached with the eigenvectors basis of S and its associated eigenvalues set.

The final point of the demonstration concerns the choice of the eigen-subspace itself. For each choice of S and for a given value of σ the maximization of the likelihood is thus equivalent to the maximization of:

$$-\frac{N}{2}\sum_{k=1}^{P}\left(\ln(\lambda_{l_k}) - \frac{\lambda_{l_k}}{\sigma^2}\right)$$

It is easy to show that the function:

$$f_\sigma(x) = \frac{x}{\sigma^2} - \ln(x)$$

is decreasing for $x < \sigma^2$ and increasing for $x > \sigma^2$. So, if we suppose that $\sigma^2 < \lambda_P$ and if we impose to the local density model the condition:

$$\mu_k > \sigma^2 \qquad \forall k, k = 1, ..., P$$

The maximum of the sum of terms within the increasing interval of the function will be reached with the set of the P highest eigenvalues and eigenvectors.

Thus, under the condition for the square of the width parameter to be smaller than the P-th highest eigenvalue of the covariance matrix, The model is optimal in term of maximum likelihood among the subspaces generated by the eigenvectors with eigenvalue greater than σ^2. In particular if $\sigma \to 0$, the model might become optimal among all the data space.

4 Training of the Mixture.

4.1 Standard EM Algorithm.

Once proved the optimality of the P largest eigenvalues for the local density, we now enter the framework of the Expectation-Maximization algorithm for the mixture case [1], [16]. The principles of the EM iterative algorithm are well known and we only recall here the main characteristics. At each time step t given the parameters set $\Theta^{(t)}$, The E-step evaluates for each pattern $x_i, i = 1, 2, ..., N$ of the learning set and each local unit $j = 1, 2, ..., M$ the term:

$$h_i^{j(t)} = (\alpha_j^{(t)} g(x_i|\Theta_j^{(t)}))/\sum_{l=1}^{M} \alpha_l^{(t)} g(x_i|\Theta_l^{(t)})$$

Then the M-step of the algorithm maximizes the expectation of the likelihood:

$$Q(\Theta, \Theta^{(t)}) = \sum_{i=1}^{N}\sum_{j=1}^{M} h_i^{j(t)} L(\Theta_j|X)$$

The mixing coefficients and the centres are classically modified by:

$$\alpha_j^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} h_i^{j(t)} \qquad u_j^{(t+1)} = \sum_{i=1}^{N} h_i^{j(t)} x_i / \sum_{i=1}^{N} h_i^{j(t)}$$

and, as in the single case, the solutions are the P largest eigenvalues of the updated local covariance matrix:

$$\Sigma_j^{(t+1)} = \sum_{i=1}^{N} h_i^{j(t)} (x_i - u_j^{(t+1)})^2 / \sum_{i=1}^{N} h_i^{j(t)}$$

Nevertheless, in practice, this EM formulation of the learning is not appealing because at each M-step we have to evaluate from scratch the P eigenvalues for each new covariance matrix. In fact we

are interested in finding a more progressive evaluation of the eigenvalues and eigenvectors we present now.

4.2 Iterative Learning.

In the EM approach for mixture of gaussians, different levels of progressive learning can be identified from an update at the M-step of time (t+1) of the eigenvalues resulting from the previous M-step to a complete on-line version with an update of all parameters [10] after each pattern. In the current work, we have tested the intermediate case by replacing the complete extraction of eigenvectors from each local covariance matrix by an iterative PCA algorithm. Starting from the values at the end of the previous M-step, this latter updates the eigenvectors in an inner loop of the M-step. The iterative method is here a local version of the Generalized Hebbian Algorithm (GHA)[13]. For the eigenvectors the algorithm is the following:

$$\tilde{V}_{jk}^{(t)(0)} = V_{jk}^{(t)}$$

$$\tilde{V}_{jk}^{(t)(i+1)} = \tilde{V}_{jk}^{(t)(i)} + \eta(t) h_i^{j(t)} (y_{ijk}) \left((x_i - u_j^{(t+1)}) - \sum_{l=0}^{k} (y_{ijl}) \tilde{V}_{jl}^{(t)(i)} \right) \qquad V_{jk}^{(t+1)} = \tilde{V}_{jk}^{(t)(N-1)}$$

with: $y_{ijk} = {}^{T}(x_i - u_j^{(t+1)}) \tilde{V}_{jk}^{(t)(i)} \quad \forall i, i = 1, \ldots, N \quad \forall j, j = 1, \ldots, M \quad \forall k, k = 1, \ldots, P$

and $\eta(t)$ a slowly time-varying parameter.We update the eigenvalues in the same loop:

$$\tilde{\lambda}_{jk}^{(t)(0)} = 0 \qquad \tilde{\lambda}_{jk}^{(t)(i+1)} = \tilde{\lambda}_{jk}^{(t)(i)} + y_{ijk}^{2} \qquad \lambda_{jk}^{(t+1)} = \tilde{\lambda}_{jk}^{(t)(N-1)}$$

In the case of the local density (M=1) this algorithm is precisely the GHA algorithm which extracts the principal components by successively decorrelating the projections of the data onto the vectors. The GHA [13] algorithm is an extension of the largest component extraction method proposed in [11]. Several proofs of convergence have been established for these algorithms [11], [13] but the detailed analysis of this convergence is still an active field of research [17]. In the case of several gaussians (M>1), the demonstration of the convergence of the algorithm towards the P principal components of each local density is not direct. Its analysis will be an interesting challenge especially on the basis of recent works [5],[16], [17].

4.3 Global Learning of the Width Parameter.

We have previously explained the choice of an unique width parameter σ for the mixture so as to be able to easily compare the influence of the different local densities. Especially in high dimension data spaces, this condition prevents the model to reduce to a Winner Take All algorithm. In this chapter, we consider the global learning of this parameter directly through the maximization of the likelihood at M-step. We first recall that we have previously shown the optimality of the principal eigenvectors and eigenvalues for any value of σ respecting the condition:

$$\sigma^2 < min(\lambda_{1P}, \lambda_{2P}, \ldots, \lambda_{MP})$$

and then the analysis of the maximization of the likelihood at the M-step with respect to σ reduces to the analysis of:

$$Q(\Theta, \Theta^{(t)}) = -\frac{N}{2}((d-P)\ln(\sigma^2) + \ln(2\Pi) + P) - \frac{1}{2\sigma^2} \sum_{j=1}^{M} \sum_{i=1}^{N} h_i^j d(x_i, u_j)^2$$

$$-\frac{N}{2} \sum_{j=1}^{M} \alpha_j \sum_{k=1}^{P} \ln(\lambda_{jk}) + \frac{N}{2\sigma^2} \sum_{j=1}^{M} \alpha_j \sum_{k=1}^{P} \lambda_{jk}$$

For the sake of simplicity we have not indicated the time step t notation. The second term of the right part of this equality can simply be expressed with the trace of the local covariance matrix:

$$\frac{1}{N} \sum_{i=1}^{N} h_i^j d(x_i, u_j)^2 = \alpha_j^{(t+1)} tr(\Sigma_j) = \alpha_j \sum_{k=1}^{d} \lambda_{jk}$$

and we thus write:

$$\frac{\partial}{\partial \sigma} Q(\Theta, \Theta^{(t)}) = -\frac{N(d-P)}{\sigma} + \frac{N}{\sigma^3} \sum_{j=1}^{M} \alpha_j^{(t+1)} \sum_{k=P+1}^{d} \lambda_{l_k}^{(t+1)}$$

The derivate of the likelihood will vanish at the value σ_0 :

$$\sigma_0^{\,2} = \frac{1}{(d-P)} \sum_{j=1}^{M} \alpha_j^{(t+1)} \sum_{k=P+1}^{d} \lambda_{jk}^{(t+1)}$$

with positive value for $\sigma^2 < \sigma_0^{\,2}$ and negative for $\sigma^2 > \sigma_0^2$ The likelihood thus presents a maximum at this value. In this condition the maximum is reached for:

$$\sigma_{opt}^{\,2} = min(\sigma_0^{\,2}, \lambda_{1P}, \lambda_{2P}, \ldots, \lambda_{MP})$$

Note that the sum of the lowest eigenvalues in the expression of σ_0^2 are not computed directly but through the trace of the covariance matrix equal to the sum of the local distances.

5 Related Mixture Approaches.

The initial motivation for the M-PCG model was to try to numerically approximate the multidimensional gaussian model by regularizing the lowest eigenvalues in a fully probabilistic context. We have therefore conducted our work from a direct eigenvalues decomposition and approximation point of view but after preparing the manuscript of [8], we have been aware of parallel works based on a Factorial Analysis approach [3], [15]. In this chapter we propose to compare our model with these very recent works.

Factorial Analysis (FA) is, beside PCA, a common tool for dimensionality reduction. In this approach, the data x of R^d is modeled by a latent variables vector y from a subspace of dimension P in the following way:

$$x = Wy + \varepsilon$$

where the vector of latent variables y is assumed to have a identity covariance distribution and the noise ε to have a diagonal covariance distribution Ψ. Therefore, the FA model implies that data are independent given the latent variables. The dxP matrix W is called the factor loading matrix and it is easily shown given these assumptions that the covariance of the data is modeled by:

$$\Sigma = \Psi + WW^T$$

Thus, like classical PCA, the FA model offers a framework for dimensionality reduction with the latent variable subspace with, unlike PCA, the great advantage of a probabilistic formulation. A common training of this density model is an iterative method is also based on the EM algorithm according to the MLE criteria [13]. At each E-step the responsibility of each latent variables for each observation is estimated, while at the M- step likelihood is maximized and factor loading and noises matrices updated. For the mixture case, this EM approach has been recently extended in a iterative way [3] but, since the estimation of each factor analyzer is based on a MLE criteria standard EM can also be applied [4].

The strong assumption of FA, expressed by the diagonality of the covariance matrix of the noise, is that data are independent given the latent variables. In practice, this is not always true as illustrated in [4] where this approach is used for modeling handwritten digit forms. In our model there is not such an hypothesis, the direct approach does not implies any kind of dependency between the components of the input variables. Moreover in classical FA the relationship between latent variables subspace and principal components are not always established. This is one of the motivation of the work presented in [15], where it is precisely shown that, in the special case where the covariance matrix of the noise reduce to $\Psi = \sigma^2 I$ the latent variables subspace corresponds, under some conditions, to the principal eigen-subspace. The local model is called PPCA and in term of dynamic is thus identical to our model PCG.

In [15], the study of the optimality of the PPCA model is very complete and leads with a specific demonstration of the optimality of the eigen-subspace for a particular value of the width parameter. Nevertheless, this demonstration is done in the FA framework that is under the assumption of independence between input variables. Another difference between the PPCA and M-PCG models lies in the use of the width parameter which in [15] is specific to each local density function and equals the mean of the lower eigenvalues of each local covariance matrix. Without anticipating about the practical characteristics of the different alternatives in term of generalization in high-dimension data spaces, the PCG local density approach we have presented is for this parameter more general since the optimality has been demonstrated in an range value of this parameter. We have also proposed a global learning for this parameter which includes the specific learning of [15].

The final difference between M-PCG and mixtures of PPCA approaches lies in the learning algorithm of the mixture. The PPCA benefits of two methods for learning the mixture. The first is the

classical extraction of a limited number of eigenvalues from each covariance matrix, which may be costly in high dimensional data space. The second is based on the FA approach for which most of the computing time is devoted to the inversion at each step of a *dxd* matrix. It is thus less costly, but still remains under the FA assumptions on the independence between the input variables. Moreover compared to the iterative GHA algorithm we have proposed, this FA approach is less progressive and for example can not be directly extended to a on-line version with an update after each pattern.

6 Numerical Experiments.

After testing the M-PCG learning algorithm on various small data sets, we have applied the method to a set of images of handwritten digits made of 2000 elements for each class. A test data set of 46634 elements is also used.With all type of data sets we have verified the convergence of the method towards local principal eigenvalues and eigenvectors for adequate values of σ. This control is easily done by applying a classical method of PCA at the end of the algorithm.

For the classification experiments, we have built for each class a mixture of five centres (M=5) and ten principal components (P=10). The total number of centres is thus 50.The first case (M-PCG-I) corresponds to the learning of the width parameter for each mixture we have presented. In the second case (M-PCG-II) an unique width parameter has been learned for all the 50 local density probabilities. In fact, this corresponds to the maximization of a joint pattern class probability likelihood of a model made of all the mixtures but with a fixed binary bayesian probability of each class for each centre. In particular, we still have the nice property of separate computing of each pattern with its own class sub-mixture during the learning phase. The third case (M-PCG-III) is an example of an experimental value of the width fixed here to 0.5 while in the second case the optimal value found was of 0.21.

The results in term of recognition rate are presented in the table 1 together with the results for the nearest neighbor method which as a costly but simple and universal method roughly illustrates the complexity of the problem to solve. The use and the global learning of an unique width parameter

	learning	test
1-nearest neighbour	100.0	94.62
M-PCG-I	95.14	94.14
M-PCG-II (σ_{opt}=0.21)	96.36	95.05
M-PCG-III (σ=0.5)	96.06	95.12

Table 1: recognition results.

in M-PCG-II appears as a better alternative than its global but class by class learning in M-PCG-I. The first reason is certainly the joint probability maximization of the width parameter in M-PCG-II which explains the better results on the learning set, but we also believe that in the data space of size 256, the classification is, due to the normalization term, excessively driven by the difference between the default width parameters for each class. The third case, with the best result on the test set illustrates the fact that the best value of the global width parameter may not be necessarily captured by the algorithm and that general or application specific regularization criteria should be applied. Here again, an advantage of the M-PCG model is to permit such a scheme since the optimality of the local density has been shown for a range value of the width parameter.

Unfortunately, the comparison with the results presented in [4] and [15] for digits classification is not direct since databases are different as well as the number of centres for each class. Future experiments will try to compare the methods on an unique database. This comparison will be certainly interesting since our model as well as the models in [4] and [15] perform slightly better than the reference nearest neighbor method [4]. We can also note that, in [4] some numerical recipes have to be used so as to prevent the absence of noise in some pixels. This is a classic way of regularizing multidimensional gaussians model but an advantage of our model is precisely to integrate this regularization aspect in its own formalism.

7 Conclusion.

We have presented in this paper the M-PCG model which is an adaptation of the mixture of gaussians model for the case of high dimensional data. This model is mainly motivated by the need for a nu-

merically tractable approximation of the mixture of gaussians model in high dimensional data spaces. It is characterized by a cheap computing time for learning and recognizing together with a generalization ability while keeping a probabilistic formulation for the local density model and a close approximation of the mixture of gaussians model. The model simply consists in replacing the lower eigenvalues of the covariance matrix by a default width value. This local density model has first been justified in term of maximum likelihood estimation. We have next proposed a learning iterative algorithm for this model which consists in replacing the M-step of the EM algorithm by an iterative method for principal components extraction based on the Generalized Hebbian Algorithm (GHA). We have also proposed a global learning for the default width parameter and focused the attractive numerical properties of our model.

The model have also been compared in theoretical terms with works based on mixture of Factorial Analyzers (FA). Especially a very similar model has emerged very recently from FA but under a theoretical aspect, it remains under the FA assumption of an independent noise between input variables. In algorithmic terms the two approaches differ in the use of the default width parameter, the iterative learning algorithm we have proposed and different numerical costs. We have finally presented the first experiments on the recognizing of handwritten digits. We have reported classification results better than the ones obtained with the 1-nearest neighbor method on our specific database and have emphasized on the interest of being able to control the default width parameter.

Under a theoretical aspect, future works will concern the establishment of a proof of convergence and its analysis. In algorithmic terms, other iterative algorithms than GHA for could be tested. Also numerical comparison with close models based on Factorial Analysis remains to be done. Nevertheless, in view of the first experiments presented here, we are yet considering the application of the model to HMM modeling for handwritten word recognition. This will be an interesting challenge for testing the ability of the model to learn in an evolving data context.

8 References.

[1] Dempster, A. P., Laird N. M. and Rubin, D. B. (1977), "Maximum Likelihood from incomplete data via the EM algorithm", J. Royal Stat. Soc. B45(1), 47-50.

[2] Fukunaga K., (1990), "statistical pattern recognition", second edition, Academic Press.

[3] Ghahramani Z. , Hinton G. E. (1997), "The EM algorithm for mixture of factorial analyzers", University of Toronto CRG-TR-96-1. ftp://ftp.cs.toronto.edu/pub/zoubin/tr-96-1.ps.gz

[4] Hinton G. E., Dayan P., Revow M., (1997), "Modelling the Manifolds of Images of Handwritten Digits", IEEE Transaction in neural networks, vol. 8, no 1, p65-74.

[5] Jordan M. I., Xu L. (1996), "Convergences Results for the EM approach to Mixtures of Experts Architextures", Neural Networks, Vol.8, no 9, 1409-1431.

[6] Kambhatla N., Leen T. K.,, (1995), " Classifying with Gaussian Mixtures and clusters", in Advances in Neural Information Processing Systems 7.

[7] Lemarié B., Gilloux M. and Leroux M.(1996), "Handwritten Word recognition using Contextual Hybrid RBF networks/Hidden Markov Models", in Advances in Neural Information Processing Systems 8.

[8] Lemarié B.,.(1998), "Mélanges de gaussiennes en composantes principales ppour l'estimation de densité", in Proceedings of RFIA, AFCET,France, ftp://ftp.srt-poste.fr/pub/rmo/bernard/rfia98.ps.Z.

[9] Mao J., Jain A. K. (1995), Artificial neural networks for feature extraction and multivariate data projection", IEE transactions in neural networks, vol 6, no 2, 1995.

[10] Neal, R. N., Hinton, G. E. (1993), "A new view of the EM algorithm that justifies Incremental and other variants", Univeristy of Toronto, Dept. of computer Science, preprint.

[11] Oja, E.(1982), " A simplified neuron model as a principal component analyzer", Journal of Mathematical Biology, 16, 267-273.

[12] Ormoneit D., Tresp V. (1996), "Improved Gaussian Mixture Density Estimates Using Bayesian Penalty Term and Network Averaging", in Advances in Neural Information Processing Systems 8.

[13] Rao, C. R., " Estimation and tests of signifiance in factor analysis", Psychometrika, 20, 93-11, 1955.

[14] Sanger, T. D. (1989), "Optimal unsupervised learning in a single layer linear feedforward network", Neural Networks, 2 , 459-473.

[15] Tipping M. E., Bishop M. C, (06/1997), Mixtures of probabilistic principal components analyser", Aston University NCRG Technical Report, http://neural-server.aston.ac.uk/cgi-bin

[16] Xu l., Jordan M. I., (1996), "on converge properties of the EM algorithm for gaussian mixtures", neural computation, 8,129-151.

[17] Xu, L. (1993), "Least Mean Square Error Reconstruction Principle for Self-Organizing Neural Nets", Neural Networks, 6, 627-648.