www.icgst.com

Mixture of Probabilistic PCA under Variance Preservation

Mohamed E. M. Musa
College of Computer Science and Information Technology
Khartoum, Sudan,
hafiz@sustech.edu,

Abstract

The idea of modeling data heterogeneity by a mixture of local models, and exploiting the correlation in the localized data subsets to reduce their subspace dimensionalities has been realized in many mixture models; like PCA mixture and FA mixture models. Instead of using fixed ad-hoc dimensionality for all subspaces, this paper proposes using a global preserved variance percentage value, to estimate the dimensionality that retains the given variability percentage in each subspace. We test the proposed method on classifying handwritten digit by a mixture of probabilistic PCA model, the result shows that the proposed method outperforms fixed dimensionality probabilistic PCA mixture model.

Keywords: *PCA, mixture model, EM algorithm.*

1. Introduction

In recent years, there is more interest in pattern recognition discriminative methods, such as support vector machines, than density-base methods. Discriminative methods find the boundaries between class regions without much interest in the structure of the class densities. Better recognition accuracy and less computational complexity are the main reasons behind the increasing interest. Despite these facts, density-based methods have many important aspects that discriminative methods could not offer:

1. Density-based methods are the choice of the dynamic class-plug-in systems, since we can train new class and plug it into the classification system without retraining the entire system.
2. Density-based methods have natural rejection criteria when all the densities are low.

Since the above mentioned aspects are critical for some applications, density based methods still undergo research to lever their performance and applicability. Model fitting when the underlying data have different structures in different parts of the input space, is one of these problems that need more research work. Fitting one global linear model for such data, can poorly represent the whole data. On the other hand, current global nonlinear models can be slow and inaccurate especially for high dimensional data [3, 8]. A combination of local linear models can quickly learn the structure of the data in local regions which consequently, offer faster and more accurate model fitting [7, 8, 13]. Partitioning training data set into smaller subsets may lead to curse of dimensionality problem, as a training sample subset may not be enough for estimating the required set of parameters for the submodel. On the other hand, increasing the size of training data is not possible in many situations. Interestingly, since the data points in local regions are highly similar, the data is highly correlated. Therefore, by decorrelation methods we can reduce data dimension and hence the number of parameters. In other words, we can find uncorrelated low dimensional subspaces that capture most of the data variability. Among these new local model methods that entail dimensionality reduction is the Mixture of Principal Component Analyzers (MPCA) [8, 5]. A turning study in this model history is Hinton et al. paper ”Modelling the Manifolds of Images of Handwritten Digits”, since this is the first work that used one global EM-training process in pseudo-likelihood framework [7]. All algorithms proposed before this one has two separate processes; partitioning the data space by hard clustering methods, followed by fitting a PCA for each cluster. Another major enhancement comes from Tipping & Bishop and Roweis through proposing the Probabilistic Principal Component Analyzer (PPCA). By giving a probabilistic definition for PCA the usage of the mixture model and soft clustering, to define the mixture of PPCA is straight forward. To this end, MPPCA model still suffers the following problems:

1. There is no standard method to specify the optimal number of subspaces.
2. There is no standard method for EM algorithm initialization, nor a standard method to help the algorithm escape the local maxima.

3. There is no standard method to specify the optimal dimensionality for each subspace.

As this model has many applications one global method for optimal solution may not be possible. For instance, the optimal number of subspaces and dimensionality for best classification performance may not be optimal for compression. Therefore, it is more reasonable to offer the modeler a graded manageable scheme to control subspaces complexity in a global way. In line with this philosophy, we propose using a global percentage constant, that represent the preserved variance retained by each subspaces. This method has many advantages over specifying global constant value for all subspaces dimensionality:

1. As pointed out by Meinicke and Ritter [9], data acquisition devices e.g. sensor, generally have the same noise percentage presence. Therefore, a global preserved variance value is more efficient in de-noising subspaces.

2. Subspaces with similar variability is expected to be more smoothed and therefore their density estimate is expected to be better than subspaces with different variability and same dimensionality. Moreover the danger of overfitting is also less.

2. Probabilistic Principal Component Analyzer

Tipping and Bishop [10] found a probabilistic formulation of PCA by viewing it as a latent variable problem, in which a *d*-dimensional observed data vector $\boldsymbol{y}$ can be described in terms of an *m*-dimensional latent vector,

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\mu} + \boldsymbol{w} \tag{1}$$

Where $\boldsymbol{A}$ is an $d \times m$ matrix, $\boldsymbol{\mu}$ is the data mean and $\boldsymbol{w}$ is an independent Gaussian noise with a diagonal covariance matrix $\boldsymbol{I}$. The probability of observed data vector y is:

$$p(\boldsymbol{y}) = (2\pi)^{-d/2}|\boldsymbol{C}|^{-1/2} \times \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right) \tag{2}$$

Where C is the model covariance matrix given by:

$$\boldsymbol{C} = \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} + \sigma^2\boldsymbol{I} \tag{3}$$

2.1 Subspace Dimensionality and Noise

Intuitively, the model partitions the input space into subspace of signal (principal subspace) and noise (σ^2 in Eqn. 3). As the variability in the minor $d - m$ dimensions hypothetically considered noise only, the eigenvalues that result form the data covariance matrix diagonalization, could be used to estimate the noise level. Tipping and Bishop have shown that the m.l.e. for the noise could be given by:

$$\sigma^2_{av} = \frac{1}{d-m}\sum_{k=m+1}^{d}\lambda_k,$$

where $\{\lambda_k\}_{m+1}^{d}$ are the $d - m$ small eigenvalues. Notice that we have subscripted σ^2_{av} to differentiate this value form the one that follows. In line with the conclusion of our previous paper [11], that says *the noise level could be used to determine the subspace dimensionality*, Meinicke and Ritter have given the same suggestion [9]. A PPCA model that has a hypothesized noise level and estimated dimensionality (VD-PPCA) is advantageous over the one that has a hypothesized fixed dimensionality and estimated noise (FD-PPCA). Meinicke and Ritter have shown that for a fixed noise level model, VD-PPCA, the m.l.e. of the principal subspace dimensionality could be given by:

$$\hat{m} = |\{\lambda_i : \lambda_i > \sigma^2_{fxd}, i = 1, \ldots, d\}| \tag{4}$$

where $\sigma^2_{fxd,}$ is a hypothesized constant value. In contrast with σ^2_{av}, σ^2_{fxd} is approximately equal to λ_{m+1}. With conventional PCA, the user can input a retained variance percentage (let us call it α), for which the system calculates the dimensionality that retains this variability percentage. We think this could also work for local PCA model. Moreover, by validation methods the system could calculate a suboptimal value for α from the given training data for the underlying application. This suboptimal value has special importance as it makes the determine the dimensionality autonomously.

3. Mixtures of Probabilistic Principal Component Analyzers

Thanks to the Probabilistic PCA model as it facilitates defining the Mixture of Probabilistic Principal Component Analyzers (MPPCA) as a restricted Mixture of Gaussians model (MoG) which could be trained globally in maximum likelihood framework. A mixture of Gaussians is given by the weighted sum:

$$f_k(\mathbf{y}|\boldsymbol{\theta}) = \sum_{j=1}^{k} q_j p_j(\mathbf{y};\boldsymbol{\theta}_{\mathbf{j}}) \tag{5}$$

where the *j-th* component $p_j(\mathrm{y}: \Theta_j)$ is a *d*-dimensional Gaussian density, parameterized by the mean μ_j, A*j* and σ_j , in our restricted Gaussian model, which are collectively denoted by the parameter vector Θ_j. As there is no direct method for training the mixture model, The EM algorithm estimates the model parameters iteratively, using the following set of equations:

- **E step**

$$P(j|\mathbf{y_i}) = \frac{q_j p(\mathbf{y_i}; \boldsymbol{\theta_j})}{f_k(\mathbf{y_i})} \quad (6)$$

- **M step**

$$q_j = \frac{1}{n}\sum_{i=1}^{n} P(j|\mathbf{y_i}), \quad (7)$$

$$\boldsymbol{\mu_j} = \frac{\sum_{i=1}^{n} P(j|\mathbf{y_i})\mathbf{y_i}}{\sum_{i=1}^{n} P(j|\mathbf{y_i})}, \quad (8)$$

$$\mathbf{S_j} = \frac{\sum_{i=1}^{n} P(j|\mathbf{y_i})(\mathbf{y_i}-\boldsymbol{\mu_j})(\mathbf{y_i}-\boldsymbol{\mu_j})^T}{\sum_{i=1}^{n} P(j|\mathbf{y_i})}. \quad (9)$$

Starting with some initial values for the unknown parameters vector Θ, the EM algorithm computes the Starting with some initial values for the unknown parameters vectors Θ, the EM algorithm computes the posteriori probabilities for the training data using Eqn. 6; this is known as expectation step (E-step). In the second step -known as maximization step (M-step)- the algorithm use the recently calculated posteriori probabilities and Eqns (7, 8, 9) to calculate new estimation for the parameters vector Θ. The calculation then, cycles from expectation to maximization and from maximization to expectation, until the revised estimate do not differ appreciably from the estimate obtained in the previous iteration or alternatively, until there is no significant change in the log likelihood value. More information about EM and its properties could be found in [4].

3.1 Training

Our training algorithm can be summarized as follows:

1. input D (training data set) and α (preserved variance).
2. find k hard clusters
3. for each cluster estimate the parameters vector Θ.
4. fit the model using the EM algorithm.
5. Re-fit all submodel dimensionality to preserved α from the local variability.
6. fit the model using the EM algorithm.

The algorithm starts by finding k hard clusters. There are two reasons for this step:

1. To help the EM to generalize by starting from well distributed clusters and escape local minima's;
2. To find some realistic, rather then ad hoc, starting parameter values.

For the hard clustering step, we use a recently developed Gaussian centres finding algorithm, proposed by Dasgupta [1]. Working in a reduced dimension subspace is the main reason for choosing this algorithm, without claiming that this is best initialization method for our purpose. At the beginning, the dimensionality of each subspace is determined by α and the set of points belonging to the underlying hard cluster only. Subspaces change their shapes during EM iteration (step 4) for this reason, we re-estimate subspaces dimensionality and fit with EM.

4 A Sample Application

Handwritten digit recognition is a popular classification problem that is used extensively in testing relative density classification approaches as well as discriminative approaches [7]. In this section we describe our experiments for testing the proposed algorithms on modeling handwritten digits using mixtures of PPCA.The data set used in our experiments was extracted from the well-known NIST handwritten digit database [12]. The original data set consisted of 128x128 pixel binary images. In pre-processing, these images were normalized for position, size, slant and stroke width, resulting in 16x16 pixel grey-value images [2]. Furthermore, for the experiments described in this paper PCA was used on the entire data set to reduce the number of dimensions from 256 to 64. The resulting data set was used to construct training and test sets. Using the given data set, we have trained a mixture of PPCA model for handwritten digit recognition, using our proposed method and tested its classification performance. During the training phase, only images of one class are presented to the model generator program, i.e. each digit model is built separately. Each digit id modelled by 10 subspaces. 1000 patterns (images) per class (digit) have been use for training and 1000 patterns per class have been used for testing. Experiments with PPCA mixture models and what alike in being mixture of Gaussians, e.g. Mixture of FA, showed that when the noise estimate is small i.e. σ^2 is small the model is prone to overfitting [7]. To overcome this problem we need to regularize the model by one of two methods. One method is by adding constant value to all σ^2 av. The other regularization way is by imposing a minimum allowable variance in all dimensions. These regularization methods are roughly equal. In this set of experiments we used the second one. Specifically we used 0.5 as a minimum allowable variance.

Table 1, shows the testing result for different α (preserved variance). For the purpose of comparing our proposed method with the fixed dimensionality method, we trained a fixed dimensionality model with its fixed dimensionality equal to the average dimensionality for all classes found by the proposed model. Fig. 1. summarizes these results graphically.

Class ’1’ has the least average subspace dimensionality. On the other hand, class ’3’,’5’ and ’8’ has the maximum average subspace dimensionality. This seems reasonable, as the later have more curvature in their shapes and their input space is more full of structure, while the former is semi straight line, This shows that the method is reasonable in its subspace dimensionality selection.

6 Conclusions

We designed an EM algorithm for Mixture of Probabilistic Principal Component Analyzers training. According to a global preserved variance value α, the algorithm determines for each subspace the dimensionality that retains α percentage of the local space variability. We applied the model to handwritten digit recognition. The result shows that the preserved variance is a suitable guidance for the process of searching for the optimal subspace dimensionality.

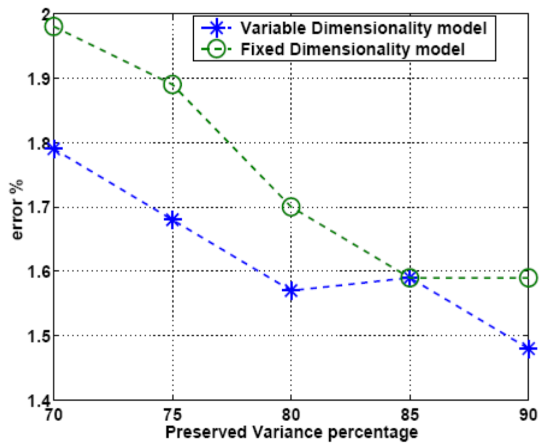

Fig. 1. Errors (% of test patterns classified incorrectly) as a function of the preserved variance percentage (α)

3. References

[1] Dasgupta, S. Learning Mixtures of Gaussians. Proc. IEEE Symposium on Foundation of Computer Science, 1999.

[2] de Ridder, D., Hoekstra, A. and Duin, R.P.W., Feature extraction in shared weights neural networks. In Kerckhoffs, E.J.H., Sloot, P.M.A., Tonino, J.F.M. and Vossepoel, A.M., editors, Proceedings of the 2nd annual conference of the Advanced School for Computing and Image Processing, pp. 289-294, Delft, The Netherlands, 1996. ASCI, ASCI.

[3] de Ridder, D.,Adaptive methods of image processing, doctoral dissertation, Faculty of Applied Science, Delft University of Technology.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Roy. Statist. Soc. B, 39:1-38, 1977.

[5] Dony, R. D., Haykin, S. Optimally adaptive transform Coding. IEEE Transactions on Image Processing, 4(10), 1358-1370.9

[6] Figueiredo, Mario A. T. and Jain, anil K. Unsupervised Learning of Finite Mixture Models. IEEE on Pattern Analysis and Machine Intelligence, 24(3):381-396,2002.

[7] Hinton, G.E., Dayan, P. and Revow, M. Modeling the manifolds of images of handwritten digits. IEEE Transaction on Neural Networks, 10(3):65-74, 1997.

[8] Kambhatla, N. Local models and Gaussian mixture models for statistical data processing. Doctoral dissertation, Oregon Graduate Institute of Science & Technology,1995

[9] Meinicke, P and H. Ritter. Local PCA Learning with Resolution-Dependent Mixtures of Gaussians Proc. International Conference on Artificial Neural Networks (ICANN'99), pages 497-502, Edinbrgh, UK, 1999.

[10] Tipping, M. E. and Bishop, C. M. Mixtures of Principal Component Analyzers, Neural Computation, 11(2):443-482,1999

[11] Musa, M. E. M., Robert P.W. Duin and Dick de Fidder. An Enhanced EM Algorithm for Mixture of Probabilistic Principal Component Analysis, ICANN 2001

[12] Wilson, C.L. and Garris, M.D. Handprinted character database 3, February 1992. National Institute of Standards and Technology; Advanced Systems Division. URL: http://www.nist.gov/srd/nistsd19.htm.

Classes	0	1	2	3	4	5	6	7	8	9	Avg.
Avg. Dim	7	4	8	9	7	8	6	5	8	6	7
V. Dim. errors	1.2	1.5	1.3	1.9	1.9	2.9	1.2	1.8	2.1	2.7	1.79
F. Dim. errors	0.9	1.2	1.7	2.8	2.3	2.8	1.4	2.3	3.5	2.6	1.98
Preserved variance .70											
Avg. Dim	9	5	9	10	8	10	8	6	10	7	8
V. Dim. errors	0.6	1.6	1.3	1.7	2.1	2.6	0.8	1.9	1.8	2.1	1.63
F. Dim. errors	0.9	1.3	1.2	2.6	3.1	3	1.2	1.5	2.6	2.6	1.89
Preserved variance .75											
Avg. Dim	11	6	11	12	10	12	9	8	12	9	10
V. Dim. errors	0.6	1.7	0.9	1.6	1.4	1.7	0.7	1.8	2.4	2.5	1.57
F. Dim. errors	0.8	1.2	1.7	2	1.9	2.9	0.6	1.8	2.4	2.2	1.7
Preserved variance .80											
Avg. Dim	14	8	14	16	13	15	12	10	15	12	13
V. Dim. errors	0.8	1.6	1	1.5	1.5	1.7	0.8	2.2	2.4	2.2	1.59
F. Dim. errors	0.8	1.2	0.9	2	2	1.9	0.9	1.8	2.8	2.6	1.59
Preserved variance .85											
Avg. Dim	18	10	18	20	16	20	15	14	19	15	17
V. Dim. errors	0.9	1.1	1.1	1.7	1.6	2.2	0.9	2	2.7	2.4	1.48
F. Dim. errors	0.8	1	1	1.5	1.8	1.8	0.6	1.9	2.4	2.5	1.50
Preserved variance .90											

Table 1: This table summarizes the classification results for different preserved variance percentage i.e. α The dimensionality of F. Dim. Models are equal to the average of the dimensionalities found by V. Dim model for all model