LOCAL PCA LEARNING WITH RESOLUTION-DEPENDENT MIXTURES OF GAUSSIANS

Peter Meinicke and Helge Ritter

Faculty of Technology, University of Bielefeld
33501 Bielefeld, Germany
Email: {pmeinick, helge}@techfak.uni-bielefeld.de

Abstract

A globally linear model, as implied by conventional Principal Component Analysis (PCA), may be insufficient to represent multivariate data in many situations. It has been known for some time that a combination of several "local" PCA's can provide a suitable approach in such cases [1, 2]. An important question is then how to find an appropriate partitioning of the data space together with a proper choice of the local numbers of principal components (PC's). In this contribution we address both problems within a density estimation framework and propose a probabilistic approach which is based on a mixture of subspace-constrained Gaussians. Thereby the number of local PC's depends on a global resolution parameter, which represents the assumed noise level and determines the degree of smoothing imposed by the model. As a consequence the model leads to an automatic resolution-dependent adjustment of the optimal principal subspace dimensionalities, which may vary among the different mixture components. Furthermore it allows to provide the optimization with an annealing scheme, which solves the initialization problem and offers an incremental model refinement procedure. Experimental results on synthetic and high-dimensional real-world data illustrate the merits of the proposed approach.

Keywords

unsupervised learning, local PCA, mixtures of Gaussians, deterministic annealing

1 Introduction

To overcome the limitations of a globally linear model, local PCA's can provide an effective means to deal with non-linear structures in multivariate data. With that type of unsupervised learning an adequate partitioning, which assigns the data to the different PCA's, is crucial for the success of that method. Several attempts [3, 4] treat the partitioning problem independently from the local PCA fitting and arrive at a somewhat arbitrary algorithmic solution without an overall optimality criterion.

Recently Tipping and Bishop [5] showed how local PCA learning can be incorporated into a Gaussian mixture modelling framework with an EM-type learning rule. Their approach is based on a linear Gaussian model, which decomposes the input space into subspaces of signal and noise. Thereby the dimensionality of the signal subspace, i.e. the number of principal components (PC's) retained by the model, is fixed and the isotropic noise variance is estimated from the data.

In this contribution we propose a complementary mixture approach, considering the noise variance as an adjustable parameter from which the soft partitioning and the subspace dimensionalities are then derived by maximum likelihood estimation. In comparison with the noise-estimating approach of [5] our dimensionality-estimating scheme offers some new and favourable features:

- It allows for different numbers of local PC's to be automatically adjusted according to some global resolution parameter, which represents the assumed noise level
- it can be used with arbitrarily high dimensional input spaces (only limited by computational resources)
- as a natural extension to maximum entropy clustering via deterministic annealing it overcomes the problem of an adequate initialization of the parameters and allows for an incremental model refinement procedure.

In general the probabilistic approach to local PCA learning can be viewed as an alternative to previous geometrically motivated approaches, which propose to fit a set of linear manifolds simultaneously to the data. Thereby the related methods usually seek to minimize the sum of orthogonal distances with respect to these manifolds, which can be defined by the

local means (centroids) and the eigenvectors of the local covariance matrices of some "nearest manifold" partitions [1, 2, 6]. In analogy to K-means hard clustering a modified version of the Generalized Lloyd algorithm can be applied to minimize the overall error. As already noted by Bezdek et. al. [1] such a procedure can lead to non-convex and often highly unplausible partitions connecting distant well separated data regions. Obviously this problem is due to the infinite level surfaces, i.e. the surfaces of constant orthogonal distance with respect to a single linear manifold. As an alternative it was proposed to use a convex combination of the orthogonal distance and the point-to-point centroid distance, whereby the resulting individual level surfaces are hyperellipsoids and hence of finite extent. Within the Gaussian framework the level surfaces (here surfaces of constant Mahalanobis distance) are also hyperellipsoids and thus the subspace-constrained Gaussian Mixture model can be viewed as a principled approach towards a well-behaved PCA partitioning of the data.

2 Resolution-Dependent Gaussian

In order to restrict the complexity of a d-variate Gaussian model with density

$$g(\mathbf{x} \mid \phi) \propto \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{1}$$

where $\phi = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we propose to use the covariance model

$$\boldsymbol{\Sigma} = \boldsymbol{\Psi} + \sigma^2 \mathbf{I}, \qquad \text{rank } \boldsymbol{\Psi} = q \tag{2}$$

with σ^2 fixed, where $\mathbf{I}$ is the $d \times d$ identity matrix and where we assume the $q \leq d$ nonzero eigenvalues of $\boldsymbol{\Psi}$ to be distinct. Thus the random vector $\mathbf{x}$ which is assumed to generate the data set $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^d$ can be interpreted to be composed of a signal part which varies in a q dimensional subspace and an isotropic Gaussian noise part with known variance. With the assumed noise variance one selects the degree of smoothing and thus the level of resolution, imposed by the model. With decreasing σ^2 the resolution increases and more subspace structure within the data becomes "visible". This fact is reflected by an increasing maximum likelihood estimator (m.l.e.) of q, the number of PC's retained by the model. Its estimation requires the calculation of the sample mean $\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$, which is the m.l.e. of $\boldsymbol{\mu}$, and the sample covariance matrix

$$\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \tag{3}$$

with eigendecomposition

$$\mathbf{S} = \mathbf{C}\,\text{diag}(l_1, \ldots, l_d)\,\mathbf{C}^T \tag{4}$$

where $l_1 > l_2 > \cdots > l_d$ are the sample eigenvalues and $\mathbf{C}$ contains the corresponding eigenvectors as columns. Then the m.l.e. of the subspace dimensionality is

$$\hat{q} = |\{l_i : l_i > \sigma^2,\ i = 1, \ldots, d\}| \tag{5}$$

which is the number of sample eigenvalues exceeding the noise variance (see [7] for a proof). Finally the m.l.e. of $\boldsymbol{\Psi}$ is

$$\hat{\boldsymbol{\Psi}} = \mathbf{C}_{\hat{q}}\,[\text{diag}(l_1, \ldots, l_{\hat{q}}) - \sigma^2 \mathbf{I}_{\hat{q}}]\,\mathbf{C}_{\hat{q}}^T \tag{6}$$

where $\mathbf{C}_{\hat{q}}$ contains the eigenvectors of the $\hat{q}$ largest eigenvalues as columns [7]. Thus a PCA is performed which neglects all PC's with variance smaller than or equal to the noise variance. If we choose $\sigma^2 \geq l_1$, i.e. the noise variance exceeds or equals the largest sample eigenvalue, then the Gaussian model is spherical with $\hat{q} = 0$. While decreasing σ^2, $\hat{q}$ increases stepwise and the model covariance is estimated within the principal subspace spanned by the first $\hat{q}$ eigenvectors of $\mathbf{S}$.

From the role of the noise variance we have a clear difference to conventional PCA: while the latter normally aims at a *small* residual variance within the $d - q$ dimensional orthogonal subspace, the above probabilistic model fits well if the residual (co)variance is *isotropic*. Nevertheless the probabilistic scheme offers a proper density model and can be used as a profitable alternative to PCA, if there is evidence that the sample is from a multivariate normal distribution.

In previous approaches the above covariance model (2) has been used with q fixed and σ^2 has been estimated from the data [8, 9, 5]. In that case the m.l.e. $\hat{\sigma}^2$ is given by the arithmetic mean of the $d - q$ smallest eigenvalues of $\mathbf{S}$ [10]. In comparison with the noise-estimating approach there are several gains that are associated with the proposed dimensionality-estimating scheme:

- in high-dimensional spaces estimation of σ^2 becomes unreliable unless a huge sample is available; in that case the classical method will usually suffer from underestimation of the noise variance, which is obvious for (nearly) singular $\mathbf{S}$

- in several applications, e.g. in pattern recognition, which require the training of different class models, with all data from the same sensor, a common noise variance seems plausible and is easily realized by the above model (see [7] for an example)
- w.r.t. a mixture model it seems plausible to allow for different local subspace dimensionalities of the component densities; within the noise-estimating framework it would be cumbersome to test all possible combinations of local dimensionalities.

Especially the last point is of particular importance, since it allows to tackle the problem of differing local numbers of PC's in a convenient way and provides some resolution-dependent insight into the dimensional structure of the data. The corresponding extension to a mixture model, with the above constrained Gaussian as component density, is straightforward and will be the subject of the next section.

3 Mixture Modelling

We will now extend the subspace-constrained Gaussian model as defined in (1) and (2) to a mixture of K such models

$$p(\mathbf{x} \mid \Phi) = \sum_{i=1}^{K} \pi_i \; g(\mathbf{x} \mid \phi_i) \tag{7}$$

where Φ comprises all parameters of the component densities $g(\mathbf{x} \mid \phi_i)$ and the mixing-weights π_i, which sum up to unity and can be interpreted as the prior probabilities that a data point is generated by the i-th component source.

3.1 EM-Optimization

To fit the above mixture model to the data, we can state the local PCA learning as an appropriate EM-optimization scheme: In order to maximize the likelihood of the resolution-dependent model given the data

$$l(\Phi \mid \mathcal{X}, \sigma^2) = \prod_{i=1}^{N} p(\mathbf{x}_i \mid \Phi) \tag{8}$$

it is sufficient to maximize the expected complete-data log-likelihood [11]

$$L(\Phi) = \sum_{i=1}^{N} \sum_{j=1}^{K} w_{ij} [\log \pi_j + \log g(\mathbf{x}_i \mid \phi_j)] \tag{9}$$

where the w_{ij} are the expected values of the binary indicator variables, which assign a data point to a certain mixture component and are normally treated as missing data. A certain w_{ij} can therefore be interpreted as the posterior probability that the i-th data point belongs to the j-th component. Given some initial values for the parameters the EM-optimization comprises two steps which are iterated until convergence:

E-Step

Calculation of the expected indicator variables

$$w_{ij} = \frac{\pi_j g(\mathbf{x}_i | \phi_j)}{p(\mathbf{x}_i | \Phi)}. \tag{10}$$

M-Step

Maximization of $L(\Phi)$, which can be achieved by solving K independent convex optimization problems, each w.r.t. the corresponding component parameters ϕ_j. Each component optimization requires calculation of the local sample mean

$$\bar{\mathbf{x}}_j = 1/n_j \sum_{i=1}^{N} w_{ij} \mathbf{x}_i \tag{11}$$

which is the m.l.e. of $\boldsymbol{\mu}_j$ with $n_j = \sum_{i=1}^{N} w_{ij}$ and the sample covariance matrix

$$\mathbf{S}_j = 1/n_j \sum_{i=1}^{N} w_{ij} (\mathbf{x}_i - \bar{\mathbf{x}}_j)(\mathbf{x}_i - \bar{\mathbf{x}}_j)^T. \tag{12}$$

Then with an eigendecomposition of $\mathbf{S}_j$ the m.l.e. of the local subspace dimensionality $\hat{q}_j$ and the corresponding $\hat{\boldsymbol{\Psi}}_j$ are analogous to (5) and (6), respectively. Finally each mixing weight is updated according to

$$\hat{\pi}_j = n_j / N. \tag{13}$$

3.2 Deterministic Annealing

Since in general every algorithm for maximizing the highly non-concave likelihood function of the mixture model relies on some initial values of the model parameters, we are faced with the problem of determining a suitable starting point for the above EM-optimization.

For that purpose consider the situation where all local means coincide in the global sample mean and the noise variance σ^2 is greater than the largest sample eigenvalue, such that all component models are spherical with equal mixing weights. This initial configuration of K identical component models is well known from maximum entropy clustering via *deterministic annealing* (DA) [12, 13], where the

noise variance acts as a "temperature" parameter which is gradually decreased during optimization. However, with all centroids coinciding in the sample mean, which is indeed the unique maximum likelihood solution for large σ^2, it doesn't make sense to increase the subspace-dimensionalities if the decreasing variance attains the largest sample eigenvalue, since all model covariance parameters would adapt identically and the mixture would behave like a single (global) subspace Gaussian.

However if we suppress the subspace adaptation and allow only the local means to change, as it is the case in DA-clustering, then if σ^2 attains the largest sample eigenvalue the initial maximum likelihood solution becomes unstable and a splitting of the centroids occurs [13], provided some small portion of noise is added to the centroid parameters. Then, while further decreasing the variance, maximum likelihood estimation w.r.t. to the centroids moves them apart and enables further splittings within remaining subgroups of coinciding means at lower "temperatures".

Thus the technique of deterministic annealing effectively solves the initialization problem by tracking the maximum likelihood solution w.r.t. to the model

$$p(\mathbf{x} \mid \boldsymbol{\theta}) \propto \sum_{i=1}^{K} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{x}-\boldsymbol{\mu}_i\|^2\right\} \tag{14}$$

with $\boldsymbol{\theta} = (\boldsymbol{\mu}_1^T, \ldots, \boldsymbol{\mu}_K^T)^T$ across increasing levels of resolution. Thus for each decrement of σ^2 the likelihood function

$$l(\boldsymbol{\theta} \mid \mathcal{X}, \sigma^2) = \prod_{i=1}^{N} p(\mathbf{x}_i \mid \boldsymbol{\theta}) \tag{15}$$

is maximized, which can again be achieved via EM-optimization. After all centroids have split up, it makes perfect sense to continue the annealing process within the previous mixture framework. Then, during the subsequent second annealing phase we allow the local subspace dimensionalities to increase according to the further decreasing noise variance and in such a way we easily realize an incremental model refinement which enables an efficient exploration of the model space. In addition we may also enable the optimization of the mixing weights which we had frozen at equal values during the first annealing phase. In table 1 the above two-phase annealing scheme is stated in a more algorithmic fashion. From the above discussion

---

❶ **Define** $\sigma^2_{\max}, \sigma^2_{\min}, \alpha \in]0,1[$

❷ **Set** $\boldsymbol{\theta} = (\bar{\mathbf{x}}^T, \ldots, \bar{\mathbf{x}}^T)^T$, $m = 0$, $\sigma_m^2 = \sigma^2_{\max}$

❸ **Maximize** $l(\boldsymbol{\theta} \mid \mathcal{X}, \sigma_m^2)$

❹ **Set** $m = m+1$, $\sigma_m^2 = \alpha\, \sigma_{m-1}^2$

❺ **If** $K_{\text{eff}} < K$ **Goto** ❸

❻ **Maximize** $l(\boldsymbol{\Phi} \mid \mathcal{X}, \sigma_m^2)$

❼ **Set** $m = m+1$, $\sigma_m^2 = \alpha\, \sigma_{m-1}^2$

❽ **If** $\sigma_m^2 > \sigma^2_{\min}$ **Goto** ❻ **Else Stop**.

Table 1: Two phase annealing scheme for incremental model refinement

an appropriate value for the initial $\sigma^2_{\max}$ is the largest eigenvalue of the sample covariance matrix. The integer K_{eff} denotes the number of numerically different means and is used to determine whether all centroids had split up. Suitable values for the variance decay rate α, which yield a tight tracking, turn out to lie typically between 0.9 and 1.0. Smaller values are possible but usually require more iterations within the single EM-optimizations. The more critical value of $\sigma^2_{\min}$ normally would have to be determined by some validation method.

4 Experimental Results

For an illustration of the performance we tested the proposed model on different data sets. In the first case we generated data with a typical cluster structure while in the second case we used some high-dimensional real-world data, which consisted of images of handwritten digits. Then the model was fitted according to the proposed two-phase annealing scheme, using the EM-algorithm for successive maximization of the log-likelihood.

4.1 Simulated Mixture Data

The data were sampled from a mixture of three trivariate Gaussians with equal apriori probabilities. Training and test sets were kept small at an equal size of $N = 100$ to simulate sparse data. The three components of the generating mixture had covariance structures with standard deviations $(0.5, 0.1, 0.1)$, $(0.5, 0.3, 0.1)$ and $(0.5, 0.4, 0.3)$ along the three principal axes, respectively, to simulate different local intrinsic dimensionalities. The result-

ing three clusters were randomly rotated and placed at the corners of an equilateral triangle with two units edge length. The corresponding data-points and the fitted model are depicted in figure 1.

During annealing the temperature decay was fixed at $\alpha = 0.9$ and $\sigma^2_{\max}$ was set to the largest eigenvalue of $\mathbf{S}$. The optimization was repeated for different K's and the first annealing phase was always run until all components had split up. The second phase was run, until the performance of the model had reached a maximum. As a performance measure we used the log-likelihood of an independent "unseen" test set which we sampled from the same distribution as the equally sized training set.

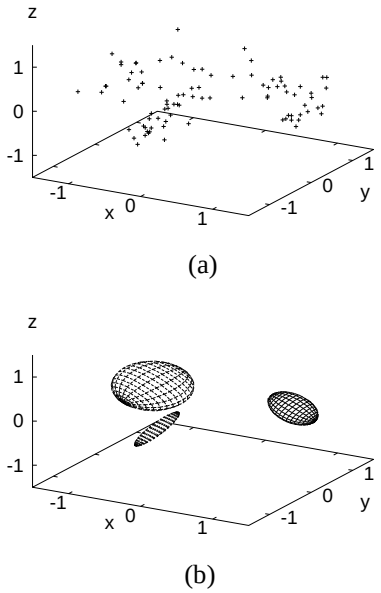

Figure 1: Data sampled from a mixture of three 3D Gaussians (a); local unit Mahalanobis-distance ellipsoids of fitted model with $K = 3$ components (b)

As it can be seen in figure 2, the model reached the highest performance for $K = 3$, where the estimated subspace dimensionalities matched the "true" ones, i.e. $\{\hat{q}_i\} = \{1, 2, 3\}$, when σ^2 approached the true noise variance which here was 0.01, corresponding to a logarithm of the inverse temperature of $\log \beta \approx 3.9$. As one may expect, the increase in likelihood is considerable as compared with the $K = 2$ model. For larger K we observed a graceful degradation, which can be seen from the curve for $K = 6$, which is still close to the optimum, as well as even larger models are.

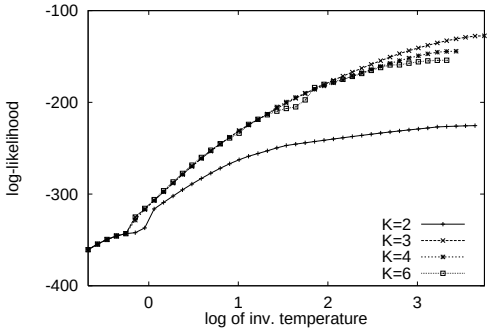

Figure 2: Performance w.r.t. to "mixture"-data; log-likelihood on test set plotted against logarithm of increasing inverse temperature $\beta = \frac{1}{2\sigma^2}$ for different numbers of mixture components $K = 2, 3, 4, 6$

4.2 Real-World Data

In particular to demonstrate the benefit of the subspace dimensionality estimation, we applied the proposed model to some high-dimensional real-world domain. The data consisted of images of handwritten digits, which were transformed to feature vectors by simply reading the pixel intensity values in lexicographic order. We used the MNIST database (http://www.research.att.com/∼yann/ocr/mnist/) which offers about 6000 training and 1000 test images per digit class. We used a 16×16 pixel format to obtain 256-dimensional data-points.

Clearly the advantage of the approach is that an optimal proportion of different numbers of local PC's is found automatically, whereas in other approaches the whole space of possible dimensionality combinations would have to be explored. As an example in figure 3 the local means and PC-directions, i.e. the retained eigenvectors of a fitted model for the digit "5" class with $K = 6$ are depicted as images. The second annealing phase had been stopped at $\sigma^2_{\min} \approx 0.1\sigma^2_{\max}$ for this example. As reported in [7] we also utilized the above formalism and the digit data for the construction of a modular plug-in classifier and achieved a competitive error-rate of 1.6 percent on the test data.

5 Conclusion

We proposed a probabilistic approach to local PCA learning which is based on a mixture of subspace-constrained Gaussians. According to some prespecified level of resolution as implied by a fixed variance noise model the formalism provides an automatic adaptation of each local number of principal components by maximum

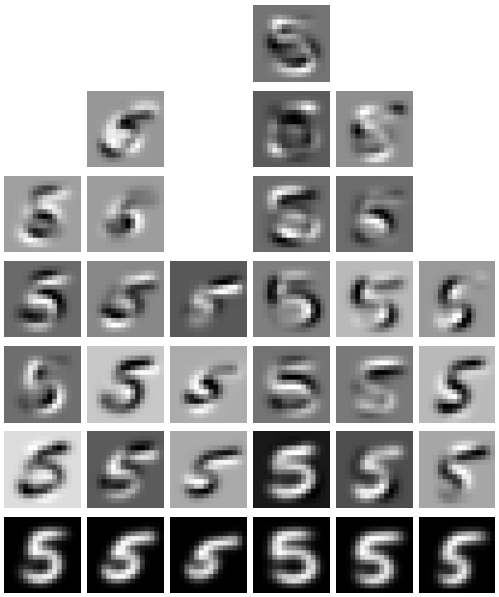

Figure 3: Local means and eigenvectors of a digit "5" model: the bottom row shows the means, the retained PC-directions are stacked on the corresponding mean with decreasing variance, i.e. the topmost "eigen-digit" accounts for least of the variance.

likelihood estimation. Together with a two-phase annealing scheme for tracking the maximum likelihood solution across increasing levels of resolution, we arrived at an incremental model refinement procedure, which avoids an improper parameter initialization and allows for an efficient exploration of the model space.

References

[1] J. C. Bezdek, C. Coray, R. Gunderson, and J. Watson. Detection and characterization of cluster substructure. *SIAM Journal on Applied Mathematics*, 40(2):339–372, 1981.

[2] Nanda Kambhaltla and Todd K. Leen. Fast non-linear dimension reduction. In *Advances in Neural Information Processing Systems*, volume 6, pages 152–159. Morgan Kaufmann Publishers, Inc., 1994.

[3] K. Fukunaga and D. R. Olsen. An algorithm for finding intrinsic dimensionality of data. *IEEE Transactions on Computers*, 20(2):176–183, 1971.

[4] Christoph Bregler and Stephen M. Omohundro. Surface learning with applications to lipreading. In Jack D. Cowan, Gerald Tesauro, and Joshua Alspector, editors, *Advances in Neural Information Processing Systems*, volume 6, pages 43–50. Morgan Kaufmann Publishers, Inc., 1994.

[5] M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analysers. *Neural Computation*, 11(2):443–482, 1999.

[6] Geoffrey E Hinton, Michael Revow, and Peter Dayan. Recognizing handwritten digits using mixtures of linear models. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7, pages 1015–1022. The MIT Press, 1995.

[7] Peter Meinicke and Helge Ritter. Resolution-based complexity control for Gaussian mixture models. Technical report, Faculty of Technology, University of Bielefeld, 1999. http://www.techfak.uni-bielefeld.de/gk/papers/.

[8] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):696–710, 1997.

[9] Sam Roweis. EM algorithms for PCA and SPCA. In Michael I. Jordan, Michael J. Kearns, and Sara A. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10. The MIT Press, 1998.

[10] T. W. Anderson. Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics*, 34:122–148, 1963.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39:1–38, 1977.

[12] R. Durbin, R. Szeliski, and A. Yuille. An analysis of the elastic net approach to the traveling salesman problem. *Neural Computation*, 1(3):348–358, 1989.

[13] K. Rose, E. Gurewitz, and G. C. Fox. Statistical mechanics and phase transitions in clustering. *Physical Review Letters*, 65(8):945–948, 1990.