The EM Algorithm for Mixtures of Factor Analyzers

Zoubin Ghahramani
Geoffrey E. Hinton
Department of Computer Science
University of Toronto
6 King's College Road
Toronto, Canada M5S 1A4

Email: zoubin@cs.toronto.edu

Technical Report CRG-TR-96-1
May 21, 1996 (revised Feb 27, 1997)

Abstract

Factor analysis, a statistical method for modeling the covariance structure of high dimensional data using a small number of latent variables, can be extended by allowing different local factor models in different regions of the input space. This results in a model which concurrently performs clustering and dimensionality reduction, and can be thought of as a reduced dimension mixture of Gaussians. We present an exact Expectation–Maximization algorithm for fitting the parameters of this mixture of factor analyzers.

1 Introduction

Clustering and dimensionality reduction have long been considered two of the fundamental problems in unsupervised learning (Duda & Hart, 1973; Chapter 6). In clustering, the goal is to group data points by similarity between their features. Conversely, in dimensionality reduction, the goal is to group (or compress) features that are highly correlated. In this paper we present an EM learning algorithm for a method which combines one of the basic forms of dimensionality reduction—factor analysis—with a basic method for clustering—the Gaussian mixture model. What results is a statistical method which concurrently performs clustering and, within each cluster, local dimensionality reduction.

Local dimensionality reduction presents several benefits over a scheme in which clustering and dimensionality reduction are performed separately. First, different features may be correlated within different clusters and thus the metric for dimensionality reduction may need to vary between different clusters. Conversely, the metric induced in dimensionality reduction may guide the process of cluster formation—i.e. different clusters may appear more separated depending on the local metric.

Recently, there has been a great deal of research on the topic of local dimensionality reduction, resulting in several variants on the basic concept with successful applications to character and face recognition (Bregler and Omohundro, 1994; Kambhatla and Leen, 1994; Sung and Poggio, 1994; Schwenk and Milgram, 1995; Hinton et al., 1995). The algorithm used by these authors for dimensionality reduction is principal components analysis (PCA).

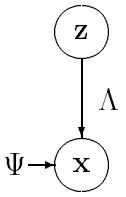

Figure 1: The factor analysis generative model (in vector form).

PCA, unlike maximum likelihood factor analysis (FA), does not define a proper density model for the data, as the cost of coding a data point is equal anywhere along the principal component subspace (i.e. the density is un-normalized along these directions). Furthermore, PCA is not robust to independent noise in the features of the data (see Hinton et al., 1996, for a comparison of PCA and FA models) . Hinton, Dayan, and Revow (1996), also exploring an application to digit recognition, were the first to extend mixtures of principal components analyzers to a mixture of factor analyzers. Their learning algorithm consisted of an outer loop of approximate EM to fit the mixture components, combined with an inner loop of gradient descent to fit each individual factor model. In this note we present an exact EM algorithm for mixtures of factor analyzers which obviates the need for an outer and inner loop. This simplifies the implementation, reduces the number of heuristic parameters (i.e. learning rates or steps of conjugate gradient descent), and can potentially result in speed-ups.

In the next section we present background material on factor analysis and the EM algorithm. This is followed by the derivation of the learning algorithm for mixture of factor analyzers in section 3. We close with a discussion in section 4.

2 Factor Analysis

In maximum likelihood factor analysis (FA), a p-dimensional real-valued data vector $\mathbf{x}$ is modeled using a k-dimensional vector of real-valued factors, $\mathbf{z}$, where k is generally much smaller than p (Everitt, 1984). The generative model is given by:

$$\mathbf{x} = \Lambda\mathbf{z} + \mathbf{u}, \tag{1}$$

where Λ is known as the *factor loading matrix* (see Figure 1). The factors $\mathbf{z}$ are assumed to be $\mathcal{N}(0, I)$ distributed (zero-mean independent normals, with unit variance). The p-dimensional random variable $\mathbf{u}$ is distributed $\mathcal{N}(0, \Psi)$, where Ψ is a diagonal matrix. The diagonality of Ψ is one of the key assumptions of factor analysis: The observed variables are independent given the factors. According to this model, $\mathbf{x}$ is therefore distributed with zero mean and covariance $\Lambda\Lambda' + \Psi$, and the goal of factor analysis is to find the Λ and Ψ that best model the covariance structure of $\mathbf{x}$. The factor variables $\mathbf{z}$ model correlations between the elements of $\mathbf{x}$, while the $\mathbf{u}$ variables account for independent noise in each element of $\mathbf{x}$.

The k factors play the same role as the principal components in PCA: They are informative projections of the data. Given Λ and Ψ, the expected value of the factors can be

computed through the linear projection:

$$E(\mathbf{z}|\mathbf{x}) = \beta\mathbf{x}, \tag{2}$$

where $\beta \equiv \Lambda'(\Psi + \Lambda\Lambda')^{-1}$, a fact that results from the joint normality of data and factors:

$$P\left(\begin{bmatrix} \mathbf{x} \\ \mathbf{z} \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Lambda\Lambda' + \Psi & \Lambda \\ \Lambda' & I \end{bmatrix}\right). \tag{3}$$

Note that since Ψ is diagonal, the $p \times p$ matrix $(\Psi + \Lambda\Lambda')$, can be efficiently inverted using the matrix inversion lemma:

$$(\Psi + \Lambda\Lambda')^{-1} = \Psi^{-1} - \Psi^{-1}\Lambda(I + \Lambda'\Psi^{-1}\Lambda)^{-1}\Lambda'\Psi^{-1},$$

where I is the $k \times k$ identity matrix. Furthermore, it is possible (and in fact necessary for EM) to compute the second moment of the factors,

$$\begin{aligned} E(\mathbf{z}\mathbf{z}'|\mathbf{x}) &= \mathrm{Var}(\mathbf{z}|\mathbf{x}) + E(\mathbf{z}|\mathbf{x})E(\mathbf{z}|\mathbf{x})' \\ &= I - \beta\Lambda + \beta\mathbf{x}\mathbf{x}'\beta', \end{aligned} \tag{4}$$

which provides a measure of uncertainty in the factors, a quantity that has no analogue in PCA.

The expectations (2) and (4) form the basis of the EM algorithm for maximum likelihood factor analysis (see Appendix A and Rubin & Thayer, 1982):

E-step: Compute $E(\mathbf{z}|\mathbf{x}_i)$ and $E(\mathbf{z}\mathbf{z}'|\mathbf{x}_i)$ for each data point $\mathbf{x}_i$, given Λ and Ψ.

M-step:

$$\Lambda^{\text{new}} = \left(\sum_{i=1}^{n} \mathbf{x}_i E(\mathbf{z}|\mathbf{x}_i)'\right)\left(\sum_{l=1}^{n} E(\mathbf{z}\mathbf{z}'|\mathbf{x}_l)\right)^{-1} \tag{5}$$

$$\Psi^{\text{new}} = \frac{1}{n} diag\left\{\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i' - \Lambda^{\text{new}} E[\mathbf{z}|\mathbf{x}_i]\mathbf{x}_i'\right\}, \tag{6}$$

where the *diag* operator sets all the off-diagonal elements of a matrix to zero.

3 Mixture of Factor Analyzers

Assume we have a mixture of m factor analyzers indexed by ω_j, $j = 1, \ldots, m$. The generative model now obeys the following mixture distribution (see Figure 2):

$$P(\mathbf{x}) = \sum_{j=1}^{m} \int P(\mathbf{x}|\mathbf{z}, \omega_j)P(\mathbf{z}|\omega_j)P(\omega_j)d\mathbf{z}. \tag{7}$$

As in regular factor analysis, the factors are all assumed to be $\mathcal{N}(0, I)$ distributed, therefore,

$$P(\mathbf{z}|\omega_j) = P(\mathbf{z}) = \mathcal{N}(0, I). \tag{8}$$

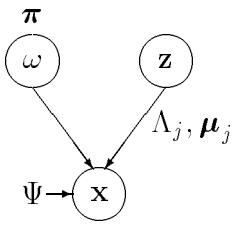

Figure 2: The mixture of factor analysis generative model.

Whereas in factor analysis the data mean was irrelevant and was subtracted before fitting the model, here we have the freedom to give each factor analyzer a different mean, $\boldsymbol{\mu}_j$, thereby allowing each to model the data covariance structure in a different part of input space,

$$P(\mathbf{x}|\mathbf{z},\omega_j) = \mathcal{N}(\boldsymbol{\mu}_j + \Lambda_j \mathbf{z}, \Psi). \tag{9}$$

The parameters of this model are $\{(\boldsymbol{\mu}_j, \Lambda_j)_{j=1}^m, \boldsymbol{\pi}, \Psi\}$; [1] the vector $\boldsymbol{\pi}$ parametrizes the adaptable mixing proportions, $\pi_j = P(\omega_j)$. The latent variables in this model are the factors $\mathbf{z}$ and the mixture indicator variable ω, where $w_j = 1$ when the data point was generated by ω_j. For the E-step of the EM algorithm, one needs to compute expectations of all the interactions of the hidden variables that appear in the log likelihood. Fortunately, the following statements can be easily verified,

$$E[w_j \mathbf{z}|\mathbf{x}_i] = E[w_j|\mathbf{x}_i]\ E[\mathbf{z}|\omega_j, \mathbf{x}_i] \tag{10}$$

$$E[w_j \mathbf{z}\mathbf{z}'|\mathbf{x}_i] = E[w_j|\mathbf{x}_i]\ E[\mathbf{z}\mathbf{z}'|\omega_j, \mathbf{x}_i]. \tag{11}$$

Defining

$$h_{ij} = E[w_j|\mathbf{x}_i] \propto P(\mathbf{x}_i, \omega_j) = \pi_j \mathcal{N}(\mathbf{x}_i - \boldsymbol{\mu}_j, \Lambda_j \Lambda_j' + \Psi) \tag{12}$$

and using equations (2) and (10) we obtain

$$E[w_j \mathbf{z}|\mathbf{x}_i] = h_{ij}\ \beta_j\ (\mathbf{x}_i - \boldsymbol{\mu}_j), \tag{13}$$

where $\beta_j \equiv \Lambda_j'(\Psi + \Lambda_j \Lambda_j')^{-1}$. Similarly, using equations (4) and (11) we obtain

$$E[w_j \mathbf{z}\mathbf{z}'|\mathbf{x}_i] = h_{ij} \left(I - \beta_j \Lambda_j + \beta_j (\mathbf{x}_i - \boldsymbol{\mu}_j)(\mathbf{x}_i - \boldsymbol{\mu}_j)' \beta_j' \right). \tag{14}$$

The EM algorithm for mixtures of factor analyzers therefore becomes:

E-step: Compute h_{ij}, $E[\mathbf{z}|\mathbf{x}_i, \omega_j]$ and $E[\mathbf{z}\mathbf{z}'|\mathbf{x}_i, \omega_j]$ for all data points i and mixture components j.

M-step: Solve a set of linear equations for π_j, Λ_j, $\boldsymbol{\mu}_j$ and Ψ (see Appendix B).

The mixture of factor analyzers is, in essence, a reduced dimensionality mixture of Gaussians. Each factor analyzer fits a Gaussian to a portion of the data, weighted by the posterior probabilities, h_{ij}. Since the covariance matrix for each Gaussian is specified through the lower dimensional factor loading matrices, the model has $mkp + p$, rather than $mp(p+1)/2$, parameters dedicated to modeling covariance structure.

[1] Note that each model can also be allowed to have a separate Ψ matrix. This, however, changes its interpretation as sensor noise.

4 Discussion

We have described an EM algorithm for fitting a mixture of factor analyzers. Matlab source code for the algorithm can be obtained from `ftp://ftp.cs.toronto.edu/pub/zoubin/mfa.tar.gz`. An extension of this architecture to time series data, in which both the factors $\mathbf{z}$ and the discrete variables ω depend on their value at a previous time step, is currently being developed.

One of the important issues not addressed in this note is model selection. In fitting a mixture of factor analyzers the modeler has two free parameters to decide: The number of factor analyzers to use (m), and the number of factor in each analyzer (k). One method by which these can be selected is cross-validation: several values of m and k are fit to the data and the log likelihood on a validation set is used to select the final values. Greedy methods based on pruning or growing the mixture may be more efficient at the cost of some performance loss. Alternatively, a full-fledged Bayesian analysis, in which these model parameters are integrated over, may also be possible.

Acknowledgements

We thank C. Bishop for comments on the manuscript. The research was funded by grants from the Canadian Natural Science and Engineering Research Council and the Ontario Information Technology Research Center. GEH is the Nesbitt-Burns fellow of the Canadian Institute for Advanced Research.

A EM for Factor Analysis

The expected log likelihood for factor analysis is

$$\begin{aligned} Q &= E\left[\log \prod_i (2\pi)^{p/2}|\Psi|^{-1/2}\exp\{-\frac{1}{2}[\mathbf{x}_i-\Lambda\mathbf{z}]'\Psi^{-1}[\mathbf{x}_i-\Lambda\mathbf{z}]\}\right] \\ &= c-\frac{n}{2}\log|\Psi| - \sum_i E\left[\frac{1}{2}\mathbf{x}_i'\Psi^{-1}\mathbf{x}_i - \mathbf{x}_i'\Psi^{-1}\Lambda\mathbf{z} + \frac{1}{2}\mathbf{z}'\Lambda'\Psi^{-1}\Lambda\mathbf{z}\right] \\ &= c-\frac{n}{2}\log|\Psi| - \sum_i \left(\frac{1}{2}\mathbf{x}_i'\Psi^{-1}\mathbf{x}_i - \mathbf{x}_i'\Psi^{-1}\Lambda\ E[\mathbf{z}|\mathbf{x}_i] + \frac{1}{2}tr\left[\Lambda'\Psi^{-1}\Lambda\ E[\mathbf{z}\mathbf{z}'|\mathbf{x}_i]\right]\right), \end{aligned}$$

where c is a constant, independent of the parameters, and tr is the trace operator.

To re-estimate the factor loading matrix we set

$$\frac{\partial Q}{\partial \Lambda} = -\sum_i \Psi^{-1}\mathbf{x}_i E[\mathbf{z}|\mathbf{x}_i]' + \sum_l \Psi^{-1}\Lambda^{\text{new}} E[\mathbf{z}\mathbf{z}'|\mathbf{x}_l] = 0$$

obtaining

$$\Lambda^{\text{new}}\left(\sum_l E[\mathbf{z}\mathbf{z}'|\mathbf{x}_l]'\right) = \sum_i \mathbf{x}_i E[\mathbf{z}|\mathbf{x}_i]'$$

from which we get equation (5).

We re-estimate the matrix Ψ through its inverse, setting

$$\frac{\partial Q}{\partial \Psi^{-1}} = \frac{n}{2}\Psi^{\text{new}} - \sum_i \left(\frac{1}{2}\mathbf{x}_i\mathbf{x}_i' - \Lambda^{\text{new}}E[\mathbf{z}|\mathbf{x}_i]\ \mathbf{x}_i' + \frac{1}{2}\Lambda^{\text{new}}E[\mathbf{z}\mathbf{z}'|\mathbf{x}_i]\Lambda^{\text{new}'}\right) = 0.$$

Substituting equation (5),

$$\frac{n}{2}\Psi^{\text{new}} = \sum_i \frac{1}{2}\mathbf{x}_i\mathbf{x}_i' - \frac{1}{2}\Lambda^{\text{new}}E[\mathbf{z}|\mathbf{x}_i]\ \mathbf{x}_i'$$

and using the diagonal constraint,

$$\Psi^{\text{new}} = \frac{1}{n}diag\left\{\sum_i \mathbf{x}_i\mathbf{x}_i' - \Lambda^{\text{new}}E[\mathbf{z}|\mathbf{x}_i]\mathbf{x}_i'\right\}.$$

B EM for Mixture of Factor Analyzers

The expected log likelihood for mixture of factor analysis is

$$Q = E\left[\log\prod_i\prod_j\left\{(2\pi)^{p/2}|\Psi|^{-1/2}\exp\{-\frac{1}{2}[\mathbf{x}_i - \boldsymbol{\mu}_j - \Lambda_j\mathbf{z}]'\Psi^{-1}[\mathbf{x}_i - \boldsymbol{\mu}_j - \Lambda_j\mathbf{z}]\}\right\}^{w_j}\right]$$

To jointly estimate the mean $\boldsymbol{\mu}_j$ and the factor loadings Λ_j it is useful to define an augmented column vector of factors

$$\tilde{\mathbf{z}} = \begin{bmatrix}\mathbf{z}\\ 1\end{bmatrix}$$

and an augmented factor loading matrix $\tilde{\Lambda}_j = [\Lambda_j\ \boldsymbol{\mu}_j]$. The expected log likelihood is then

$$\begin{aligned} Q &= E\left[\log\prod_i\prod_j\left\{(2\pi)^{p/2}|\Psi|^{-1/2}\exp\{-\frac{1}{2}[\mathbf{x}_i - \tilde{\Lambda}_j\tilde{\mathbf{z}}]'\Psi^{-1}[\mathbf{x}_i - \tilde{\Lambda}_j\tilde{\mathbf{z}}]\}\right\}^{w_j}\right] \\ &= c - \frac{n}{2}\log|\Psi| - \sum_{i,j}\frac{1}{2}h_{ij}\mathbf{x}_i'\Psi^{-1}\mathbf{x}_i - h_{ij}\mathbf{x}_i'\Psi^{-1}\tilde{\Lambda}_j\ E[\tilde{\mathbf{z}}|\mathbf{x}_i,\omega_j] + \frac{1}{2}h_{ij}tr\left[\tilde{\Lambda}_j'\Psi^{-1}\tilde{\Lambda}_j\ E[\tilde{\mathbf{z}}\tilde{\mathbf{z}}'|\mathbf{x}_i,\omega_j]\right] \end{aligned}$$

where c is a constant. To estimate $\tilde{\Lambda}_j$ we set

$$\frac{\partial Q}{\partial \tilde{\Lambda}_j} = -\sum_i h_{ij}\Psi^{-1}\mathbf{x}_i E[\tilde{\mathbf{z}}|\mathbf{x}_i,\omega_j]' + h_{ij}\Psi^{-1}\tilde{\Lambda}_j^{\text{new}}E[\tilde{\mathbf{z}}\tilde{\mathbf{z}}'|\mathbf{x}_i,\omega_j] = 0.$$

This results in a linear equation for re-estimating the means and factor loadings,

$$\left[\Lambda_j^{\text{new}}\ \boldsymbol{\mu}_j^{\text{new}}\right] = \tilde{\Lambda}_j^{\text{new}} = \left(\sum_i h_{ij}\mathbf{x}_i E[\tilde{\mathbf{z}}|\mathbf{x}_i,\omega_j]'\right)\left(\sum_l h_{lj}E[\tilde{\mathbf{z}}\tilde{\mathbf{z}}'|\mathbf{x}_l,\omega_j]\right)^{-1} \tag{15}$$

where

$$E[\tilde{\mathbf{z}}|\mathbf{x}_i, \omega_j] = \begin{bmatrix} E[\mathbf{z}|\mathbf{x}_i, \omega_j] \\ 1 \end{bmatrix}$$

and

$$E[\tilde{\mathbf{z}}\tilde{\mathbf{z}}'|\mathbf{x}_l, \omega_j] = \begin{bmatrix} E[\mathbf{z}\mathbf{z}'|\mathbf{x}_l, \omega_j] & E[\mathbf{z}|\mathbf{x}_l, \omega_j] \\ E[\mathbf{z}|\mathbf{x}_l, \omega_j]' & 1 \end{bmatrix}.$$

We re-estimate the matrix Ψ through its inverse, setting

$$\frac{\partial Q}{\partial \Psi^{-1}} = \frac{n}{2}\Psi^{\text{new}} - \sum_{ij} \frac{1}{2} h_{ij} \mathbf{x}_i \mathbf{x}_i' - h_{ij} \tilde{\Lambda}_j^{\text{new}} E[\tilde{\mathbf{z}}|\mathbf{x}_i, \omega_j] \mathbf{x}_i' + \frac{1}{2} h_{ij} \tilde{\Lambda}_j^{\text{new}} E[\tilde{\mathbf{z}}\tilde{\mathbf{z}}'|\mathbf{x}_i, \omega_j] \tilde{\Lambda}_j^{\text{new}'} = 0.$$

Substituting equation (15) for $\tilde{\Lambda}_j$ and using the diagonal constraint on Ψ we obtain,

$$\Psi^{\text{new}} = \frac{1}{n} diag \left\{ \sum_{ij} h_{ij} \left(\mathbf{x}_i - \tilde{\Lambda}_j^{\text{new}} E[\tilde{\mathbf{z}}|\mathbf{x}_i, \omega_j] \right) \mathbf{x}_i' \right\}. \tag{16}$$

Finally, to re-estimate the mixing proportions we use the definition,

$$\pi_j = P(\omega_j) = \int P(\omega_j|\mathbf{x}) P(\mathbf{x})\, d\mathbf{x}.$$

Since $h_{ij} = P(\omega_j|\mathbf{x}_i)$, using the empirical distribution of the data as an estimate of $P(\mathbf{x})$ we get

$$\pi_j^{\text{new}} = \frac{1}{n} \sum_{i=1}^{n} h_{ij}.$$

References

Bregler, C. and Omohundro, S. M. (1994). Surface learning with applications to lip-reading. In Cowan, J. D., Tesauro, G., and Alspector, J., editors, *Advances in Neural Information Processing Systems 6*, pages 43–50. Morgan Kaufman Publishers, San Francisco, CA.

Duda, R. O. and Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. Wiley, New York.

Everitt, B. S. (1984). *An Introduction to Latent Variable Models*. Chapman and Hall, London.

Hinton, G., Revow, M., and Dayan, P. (1995). Recognizing handwritten digits using mixtures of Linear models. In Tesauro, G., Touretzky, D., and Leen, T., editors, *Advances in Neural Information Processing Systems 7*, pages 1015–1022. MIT Press, Cambridge, MA.

Hinton, G. E., Dayan, P., and Revow, M. (1996). Modeling the manifolds of Images of handwritten digits. *Submitted for Publication*.

Kambhatla, N. and Leen, T. K. (1994). Fast non-linear dimension reduction. In Cowan, J. D., Tesauro, G., and Alspector, J., editors, *Advances in Neural Information Processing Systems 6*, pages 152–159. Morgan Kaufman Publishers, San Francisco, CA.

Rubin, D. and Thayer, D. (1982). EM algorithms for ML factor analysis. *Psychometrika*, 47(1):69–76.

Schwenk, H. and Milgram, M. (1995). Transformation invariant autoassociation with application to handwritten character recognition. In Tesauro, G., Touretzky, D., and Leen, T., editors, *Advances in Neural Information Processing Systems 7*, pages 991–998. MIT Press, Cambridge, MA.

Sung, K.-K. and Poggio, T. (1994). Example-based learning for view-based human face detection. *MIT AI Memo 1521, CBCL Paper 112*.