Model-based Density Estimation by Independent Factor Analysis

Daniela G. Calò, Angela Montanari, and Cinzia Viroli

Department of Statistics,
University of Bologna, Italy
{calo,montanari,viroli}@stat.unibo.it

Abstract. In this paper we propose a model based density estimation method which is rooted in Independent Factor Analysis (IFA). IFA is in fact a generative latent variable model, whose structure closely resembles the one of an ordinary factor model but which assumes that the latent variables are mutually independent and distributed according to Gaussian mixtures. From these assumptions, the possibility of modelling the observed data density as a mixture of Gaussian distributions too derives. The number of free parameters is controlled through the dimension of the latent factor space. The model is proved to be a special case of mixture of factor analyzers which is less parameterized than the original proposal by McLachlan and Peel (2000). We illustrate the use of IFA density estimation for supervised classification both on real and simulated data.

1 Introduction

Finite mixtures of distributions represent a widely used and flexible approach to model based density estimation (see, for instance, McLachlan and Peel, 2000a). For multivariate continuous data, the preferred solution is based on the use of multivariate normal components, because of their computational convenience. This approach is usually named Gaussian mixture modelling. In this context, the p-dimensional density of a random variable $\mathbf{x}$ is modelled as a mixture of m multivariate normal densities in some unknown proportions $w_1, \ldots, w_m$:

$$f(\mathbf{x}) = \sum_{l=1}^{m} w_l \phi(\mathbf{x}; \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l), \tag{1}$$

where $\phi(\mathbf{x}, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$ denotes the p-variate normal density function with mean $\boldsymbol{\mu}_l$ and covariance matrix $\boldsymbol{\Sigma}_l$. Here, the set of unknown parameters consists of the mixing proportions w_l, the elements of the component means $\boldsymbol{\mu}_l$ and the distinct elements of the component-covariance matrix $\boldsymbol{\Sigma}_l$ for $l = 1, ..., m$.

It is worth noting that an m-component mixture can be thought of as the density of an heterogeneous population consisting of m groups. For each observed unit an allocation variable, z, may be defined which denotes the identity of the group from which the object is drawn. More precisely, z may be thought of as a multinomial random variable consisting of 1 draw on m

categories with probabilities $w_1, \ldots, w_m$. If we assume that the vector $\mathbf{x}$ is conditionally normally distributed given z, then the unconditional density of $\mathbf{x}$ (that is, its marginal density) yields a Gaussian mixture model with m components:

$$f(\mathbf{x}) = \sum_z f(z) f(\mathbf{x}|z) = \sum_{l=1}^{m} w_l \phi(\mathbf{x}; \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l). \tag{2}$$

The Gaussian mixture model (1) can be fitted iteratively to an observed sample by maximum likelihood via the expectation-maximization (EM) algorithm (Dempster *et al.*, 1977). The number of components m can be taken sufficiently large to provide an arbitrarily accurate estimate of the underlying density function.

Model (1) with unrestricted component-covariance matrices is a highly parameterized model with a total of $m-1$ (the mixing proportions) + $m \times p$ (the mean vectors components) + $m \times \frac{p(p+1)}{2}$ (the component-covariance matrices distinct elements) parameters. As the number of components m in the mixture model increases, the total number of parameters can quickly become very large relative to the sample size n, thus leading to overfitting.

With the aim of reducing the number of parameters which must be estimated in order to fit a Gaussian mixture model, Banfield and Raftery (1993) introduced a parameterization of the generic component-covariance matrix $\boldsymbol{\Sigma}_l$ based on a variant of the standard spectral decomposition of $\boldsymbol{\Sigma}_l$, which reaches its simplest structure when the component-covariance matrices are assumed to be spherical.

A different approach has been proposed by McLachlan and Peel (2000b) who suggest to adopt a mixture of factor analyzers model.

In this paper we briefly review McLachlan and Peel's approach and present a new one based on Independent Factor Analysis (Attias, 1999), which still gives a mixture of factor analyzers but involves fewer parameters than McLachlan and Peel's solution. The proposed method is applied to real and simulated data in a supervised classification context.

2 Mixtures of Factor Analyzers

McLachlan and Peel (2000b) assume that, given a sample of n observations, the distribution of each observation can be modelled, with probability w_l ($l = 1, \ldots, m$), according to an ordinary factor analysis model as

$$\mathbf{x} = \boldsymbol{\mu}_l + \mathbf{B}_l \mathbf{y}_l + \mathbf{e}_l, \tag{3}$$

where $\mathbf{y}_l$ is a q-dimensional vector of common latent variables or factors, $\mathbf{B}_l$ is a $p \times q$ matrix of factor loadings. The $\mathbf{y}_l$ are assumed to be distributed as a $\mathcal{N}(\mathbf{0}, \mathbf{I}_q)$, independently of the errors $\mathbf{e}_l$, which are distributed as $\mathcal{N}(\mathbf{0}, \mathbf{D}_l)$, where $\mathbf{D}_l$ is a diagonal matrix.

Thus, the mixture of factor analyzers model is given by (1) where the l^{th} component-covariance matrix has the form

$$\boldsymbol{\Sigma}_l = \mathbf{B}_l \mathbf{B}_l^T + \mathbf{D}_l \tag{4}$$

The set of unknown parameters consists now of the elements of the $\boldsymbol{\mu}_l$, $(m \times p)$, the $\mathbf{B}_l$, $(m \times (p \times q))$ and the $\mathbf{D}_l$, $(m \times p)$, along with the mixing proportions w_l, $(m-1)$.

In this way the mixture of factor analyzers provides a way of controlling the number of parameters through the reduced model for the component-covariance matrices, yielding a solution which is intermediate between the independence and unrestricted models.

The mixture of factor analyzers model can be fitted by using the alternating expectation-condition maximization (AECM) algorithm (see McLachan *et al.*, 2003). A formal test for the number of factors can be performed using the likelihood ratio statistic; however, in situations when n is not large relative to the number of unknown parameters, the BIC criterion might be preferable.

3 Independent Factor Analysis

Independent Factor Analysis (IFA) has been originally developed as a latent variable model for solving the problem of *blind source separation* (Attias, 1999) but it may also be interpreted as an approach to model based density estimation. In effect, Independent Factor Analysis defines a latent variable probabilistic model for the observed multivariate data:

$$\mathbf{x} = \boldsymbol{\Lambda}\mathbf{y} + \mathbf{e}. \tag{5}$$

The mean centered p observed variables $\mathbf{x}$ are assumed to arise from a smaller set of q latent factors $\mathbf{y}$, that are mixed together by the matrix $\boldsymbol{\Lambda}$. A p-dimensional Gaussian noise term $\mathbf{e}$ with zero mean and diagonal covariance matrix $\boldsymbol{\Psi}$ is added in order to account for the intrinsic variability of the observed random vector. The factors are assumed to be mutually statistically independent (and also independent from the error terms) and to have arbitrary distributions. In order to make the model flexible enough to account for arbitrary factor densities, while being analytically tractable, each factor marginal density is modelled by a mixture of m_i univariate Gaussian components:

$$f(y_i) = \sum_{l=1}^{m_i} w_{il}\phi(y_i; \mu_{il}, \nu_{il}), \tag{6}$$

for $i = 1, \ldots, q$, where μ_{il} and ν_{il} are the mean and the variance of the unidimensional Gaussian components.

As a consequence of the independence condition and of the mixture modelling assumption, in the latent space the factor joint density takes the form

$$f(\mathbf{y}) = \prod_{i=1}^{q} f(y_i) = \prod_{i=1}^{q} \sum_{l=1}^{m_i} w_{il}\phi(y_i; \mu_{il}, \nu_{il}) = \prod_{i=1}^{q} \sum_{z_i} f(z_i) f(y_i|z_i), \qquad (7)$$

where the last part has been rephrased in terms of the q allocation variables z_i. Let $\mathbf{z} = [z_1, \ldots, z_q]^T$ be a q-variate allocation variable. Then, from (7)

$$f(\mathbf{y}) = \sum_{\mathbf{z}} \left[\prod_{i=1}^{q} f(z_i) \prod_{i=1}^{q} f(y_i|z_i) \right] = \sum_{\mathbf{z}} f(\mathbf{z}) f(\mathbf{y}|\mathbf{z}), \qquad (8)$$

where $f(\mathbf{y}|\mathbf{z}) = \phi(\boldsymbol{\mu}_\mathbf{z}, \mathbf{V}_\mathbf{z})$ and $\boldsymbol{\mu}_\mathbf{z}$ and $\mathbf{V}_\mathbf{z}$ are respectively defined as:

$$\boldsymbol{\mu}_\mathbf{z} = \left[\prod_{l=1}^{m_1} \mu_{1,l}^{z_{1,l}}, ..., \prod_{l=1}^{m_q} \mu_{q,l}^{z_{q,l}} \right] \qquad \mathbf{V}_\mathbf{z} = \text{diag} \left[\prod_{l=1}^{m_1} \nu_{1,l}^{z_{1,l}}, ..., \prod_{l=1}^{m_q} \nu_{q,l}^{z_{q,l}} \right].$$

Thus $f(\mathbf{y})$ is but a q-dimensional mixture model whose generic component density is the product of q normal densities, which is normal too.

Fitting an IFA model therefore amounts to fitting a Gaussian mixture model in a low-dimensional space; therefore, just like model (1) its fit to an observed sample can be performed by maximum likelihood via the EM algorithm.

But the aspect which is more interesting from our perspective is that in so doing it also allows to model the density of the observed variables as a Gaussian mixture model. The density of the observed random vector $\mathbf{x}$ may be derived by integrating the complete data distribution with respect to the factors and by summing with respect to all the possible states of the allocation vector $\mathbf{z}$. After some calculus, the following expression is obtained:

$$f(\mathbf{x}) = \sum_{\mathbf{z}} \int f(\mathbf{x},\mathbf{y},\mathbf{z}) d\mathbf{y} = \sum_{\mathbf{z}} \int f(\mathbf{z}) f(\mathbf{x},\mathbf{y}|\mathbf{z}) d\mathbf{y} = \sum_{\mathbf{z}} f(\mathbf{z}) f(\mathbf{x}|\mathbf{z}) \quad (9)$$

where $f(\mathbf{x}|\mathbf{z}) = \int f(\mathbf{x}|\mathbf{y},\mathbf{z}) f(\mathbf{y}|\mathbf{z}) d\mathbf{y}$ is the convolution of the following densities:

$$f(\mathbf{y}|\mathbf{z}) = \phi(\boldsymbol{\mu}_\mathbf{z}, \mathbf{V}_\mathbf{z})$$

and

$$f(\mathbf{x}|\mathbf{y},\mathbf{z}) = \phi(\boldsymbol{\Lambda}\mathbf{y}, \boldsymbol{\Psi})$$

since $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$.

Since the convolution of two Gaussian densities is still Gaussian:

$$f(\mathbf{x}|\mathbf{z}) = \phi\left(\mathbf{x}|\mathbf{z}; \boldsymbol{\Lambda}\boldsymbol{\mu}_\mathbf{z}, \boldsymbol{\Lambda}\mathbf{V}_\mathbf{z}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}\right), \qquad (10)$$

and therefore, from (9), $f(\mathbf{x})$ is a Gaussian mixture model with $\prod_{i=1}^{q} m_i = m$ components.

Equation (10) clearly shows that the IFA model yields a mixture of factor analyzers too, where the generic component-covariance matrix may be expressed as $\mathbf{B}_l\mathbf{B}_l^T + \boldsymbol{\Psi}$, with $\mathbf{B}_l = \boldsymbol{\Lambda}\mathbf{V}_{\mathbf{z}}^{1/2}$.

It is evident from this formulation that such a model gives component-covariance matrices which vary from one component to another but which involve fewer parameters than McLachlan and Peel's mixture of factor analyzers.

In fact, assuming a number q of IFA factors equal to the number of factors involved in (3) and the same number m of mixture components for both the models, the parameters needed to estimate all the Σ_l for $l = 1, ..., m$, in the IFA model are only the $p \times q$ factor loadings in $\boldsymbol{\Lambda}$, the p diagonal elements of $\boldsymbol{\Psi}$ and the $\sum_{i=1}^{q} m_i$ diagonal elements of the matrices $\mathbf{V}_{\mathbf{z}}$. It is a total of $pq + p + \sum_{i=1}^{q} m_i$ parameters which can be easily proved to be less than $(pq + p)m$ which is the number of parameters involved in McLachlan and Peel's modelling of the whole set of component-covariance matrices.

Equation (10) also shows that a further reduction in the number of free parameters regards the component mean vectors, which are constrained to lie on the q-dimensional subspace spanned by the column of $\boldsymbol{\Lambda}$.

Just like in the approach based on mixture of factor analyzers, the correct IFA model specification in terms of the optimal q can be derived by making use of information criteria.

4 Some results in supervised classification

The estimation of an unknown probability density function plays a central role in many applications of multivariate techniques. For instance, in the general classification context the goal is to define a rule for the assignment of one new unit, on which a p-variate vector of variables $\mathbf{x}$ has been observed, to the class, out of G unordered ones, from which it comes. Denoted by f_g, with $g = 1, \ldots, G$, the class conditional densities and by π_g the *a priori* probability of observing an individual from population g, the so-called Bayes decision rule suggests to allocate $\mathbf{x}$ to the population $\hat{g}$ such that

$$\hat{g} = \arg\max_{g=1,\ldots,G} \{f_g(\mathbf{x})\pi_g\} \tag{11}$$

In most applications neither $f_g(x)$ nor π_g $(g = 1, \ldots, G)$ are known. In this context the use of mixture models for density estimation represents a relevant solution, which is recently receiving increasing attention. In the following our proposed solution, based on IFA, is employed for the analysis of both simulated and real data sets and its performances are compared with those of other methods based on mixtures, including also the one by McLachlan and Peel.

4.1 Simulated data

The first case study is the popular waveform data. This example has been taken from (Breiman *et al.*, 1984) and subsequently used in many works on classification, since it is considered a difficult pattern recognition problem. It is a three class problem with 21 variables, which are defined by

$$\begin{aligned} x_i &= uh_1(i) + (1-u)h_2(i) + \varepsilon_i \quad \text{Class 1} \\ x_i &= uh_1(i) + (1-u)h_3(i) + \varepsilon_i \quad \text{Class 2} \\ x_i &= uh_2(i) + (1-u)h_3(i) + \varepsilon_i \quad \text{Class 3} \end{aligned}$$

where $i = 1, \ldots, 21$, u is uniform on [0,1], ε_i are standard normal random variables and h_1, h_1 and h_3 are the following shifted triangular forms: $h_1(i) = \max(6 - |i - 11|, 0)$, $h_2(i) = h_1(i - 4)$ and $h_3(i) = h_1(i + 4)$. The optimal error rate for this data set is 0.14. The method discussed here is compared with the following classification procedures: linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), mixture discriminant analysis (MDA), flexible discriminant analysis (FDA), penalized discriminant analysis (PDA) and the CART procedure. The training sample consists of 300 observations and the test sample has size 500. Both of them have been generated with equal priors.

Technique	Error rates	
	Training	Test
LDA	0.121(.006)	0.191(.006)
QDA	0.039(.004)	0.205(.006)
CART	0.072(.003)	0.289(.004)
FDA/MARS (degree=1)	0.100(.006)	0.191(.006)
FDA/MARS (degree=2)	0.068(.004)	0.215(.002)
MDA (3 subclasses)	0.087(.005)	0.169(.006)
MDA (3 subclasses, penalized 4df)	0.137(.006)	0.157(.005)
PDA (penalized 4df)	0.150(.005)	0.171(.005)
Factor analyzers (4 subclasses)	0.129(.010)	0.187(.005)
IFA (2 factors)	0.054(.010)	0.133(.004)

Table 1. Results for waveform data. The values are averages over 10 simulations, with the standard error of the average in parentheses. The first eight entries are taken from Hastie and Tibshirani (1996). The last line indicates the error rates in the IFA with 2 components for each factor.

Table 1 indicates the classification results taken from Hastie and Tibshirani (1996) and includes the performances of IFA over 10 simulations. IFA based discriminant analysis shows the lowest classification error rate in the test samples (which is lower than the optimal one only because of sampling error).

4.2 Real data

We also applied the proposed method on the thyroid data (Coomans *et al.*, 1983). The example consists of 5 measurements (T3-resin uptake test, Total Serum thyroxin, Total serum triiodothyronine, Basal thyroid-stimulating hormone and maximal absolute difference of TSH value after injection of 200 micro grams of thyrotropin-releasing hormone as compared to the basal value) on 215 patients, that are distinguished in three groups on the basis of their thyroid status (normal, hyper and hypo). The data have been randomly divided into a training sample of size 143 and a test sample that consists of the remaining patients. Table 2 shows a summary of the performance of several classification procedures. In order to compare our results with those published in a technical report which represents an extended version of Hastie and Tibshirani (1996), only one split into training and test set has been considered. IFA based discriminant analysis performs very well and it is competitive with respect to non linear methods such as neural networks and the MDA/FDA procedure.

Technique	Error rates	
	Training	Test
LDA	0.091	0.083
MDA	0.028	0.042
MDA/FDA	0.049	0.014
FDA	0.049	0.042
Neural network (10 hidden units)	0.000	0.027
Factor analyzers (4 subclasses)	0.028	0.069
IFA (2 factors)	0.056	0.027

Table 2. Results for Thyroid data. The first five lines are taken from an extended version (technical report) of the paper by Hastie and Tibshirani (1996). The last entry indicates the error rates in IFA with 2 components for each factor.

5 Conclusions

Density estimation based on independent factors seems to give very good results which are comparable, and sometimes better, to those obtained by using other approaches still based on mixture models, but requiring the estimate of more parameters. The most relevant limit of the procedure is that it does not allow to explore the whole range of possible mixture component number since, the number of estimated components in the $\mathbf{x}$ space is dependent on both the number q of independent factors and the number of components used in order to model each factor density.

References

ATTIAS, H., (1999): Independent Factor Analysis. *Neural Computation*, 11, 803–851.

BANFIELD, J.D. and RAFTERY, A.E., (1993): Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49, 803–821.

BREIMAN, L., FRIEDMAN, J., OLSHEN, R., and Stone, C., (1984): *Classification and Regression Trees*. Wadsworth, Belmont, California.

COOMANS, D., BROECKAERT, M. AND BROEACKAERT, D.L., (1983): Comparison of multivariate discriminant techniques for clinical data - application to the tyroid functional state. *Meth. Inform. Med.*, 22, 93–101.

DEMPSTER, N.M., LAIRD, A.P., and RUBIN, D.B., (1977): Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, 39, 1–38.

HASTIE, T. and TIBSHIRANI, R., (1996): Discriminant Analysis by Gaussian Mixtures. *Journal of the Royal Statistical Society B*, 58, 155–176.

MCLACHLAN, G.J. and PEEL, D. (2000a): *Finite Mixture Models*. John Wiley & Sons INC, New York.

MCLACHLAN, G.J. and PEEL, D. (2000b): Mixtures of Factor Analyzers. In: Langley, P. (Ed.): *Proceedings of the Seventeenth International Conference on Machine Learning*. Morgan Kaufmann , San Francisco, 599–606.

MCLACHLAN, G.J., PEEL, D., and Bean, R.W. (2003): Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics and Data Analysis*, 41, 379–388.