Independent Factor Discriminant Analysis

Angela Montanari, Daniela Giovanna Caló, and Cinzia Viroli

Statistics Department
University of Bologna,
via Belle Arti 41, 40126, Bologna, Italy
(e-mail: montanari@stat.unibo.it, calo@stat.unibo.it,
viroli@stat.unibo.it)

Abstract. In the general classification context the recourse to the so-called Bayes decision rule requires to estimate the class conditional probability density functions. In this paper we propose a mixture model for the observed variables which is derived by assuming that the data have been generated by an independent factor model. Independent factor analysis is in fact a generative latent variable model whose structure closely resembles the one of ordinary factor model but it assumes that the latent variables are mutually independent and not necessarily Gaussian. The method therefore provides a dimension reduction together with a semiparametric estimate of the class conditional probability density functions. This density approximation is plugged into the classic Bayes rule and its performance is evaluated both on real and simulated data.

Keywords: Classification, Independent Factor Analysis, Mixture Models.

1 Introduction

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. The training sample on which the rule is built consists of an indication of the class membership and of the p predictors for a set of n units. 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

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

If the class conditional densities are Gaussian, the expression (1) simply yields the well known linear or quadratic discriminant functions according to whether the condition of homoscedasticity is fulfilled or not. But in most applications neither $f_g(x)$ nor π_g $(g = 1, \ldots, G)$ are known and the recourse to the Gaussian based approach may be strongly misleading.

When the training sample data may be considered as a random sample from the pooled population, the prior probabilities may be easily estimated by the relative frequencies of the g classes in the sample $\hat{\pi}_g = n_g/n$ where n_g is the number of units from class g observed in the training sample. The estimation of the unknown densities f_g is on the contrary a more complex task.

The solution most often used in the statistical literature is based on kernel density estimation ([Hand, 1982] and [Silverman, 1986]) and on the use of the estimated densities in the classification rule (1), which therefore becomes a nonparametric one. It is well known however that kernel methods deeply suffer from the curse of dimensionality when applied in the multidimensional context (as the one we are dealing with is). They also tend to produce poor estimates of the density tails, whose role may on the contrary be crucial for classification purposes. Amato *et al.* [Amato *et al.*, 2002] suggest to overcome the problem by transforming the data into independent components [Comon, 1994]. Exploiting the independence condition they rephrase the multivariate density estimation task as a sequence of univariate ones, *i.e.* the estimation of the marginal densities, whose product yields the multivariate density in the transformed space. This density is then back-transformed in order to obtain an estimate of the probability density function of the observed variables. The method seems to outperform linear, quadratic and flexible discriminant analysis in the training set, but its performance is quite poor in the test one.

Other approaches to nonparametric density estimation for classification, such as the one due to Polzehl [Polzehl, 1995], who suggests a discrimination oriented version of projection pursuit density estimation, seem to produce quite good results but at a high computational cost and many aspects, at least from an algorithmic point of view, still need improvement (for instance, the selection of the bandwidth parameters in univariate kernel density estimation, which should be optimal from a classification perspective).

A more recent approach is based on mixture models. In particular, in [McLachlan and Peel, 2000] each class conditional density, f_g, is modeled as a mixture of m_g normally distributed components (Gaussian Mixture Model, GMM):

$$\hat{f}_g(\mathbf{x}) = \sum_{l=1}^{m_g} w_{gl}\phi(\mathbf{x}; \boldsymbol{\mu}_{gl}, \boldsymbol{\Sigma}_{gl}) \qquad (2)$$

where $\phi(\mathbf{x}, \boldsymbol{\mu}_{gl}, \boldsymbol{\Sigma}_{gl})$ denotes the p-variate normal density function with vector mean $\boldsymbol{\mu}_{gl}$ and covariance matrix $\boldsymbol{\Sigma}_{gl}$ $(l = 1, \ldots, m_g)$, and w_{gl} are the mixing proportions. The density estimation involves therefore the estimation of $\boldsymbol{\mu}_{gl}$, $\boldsymbol{\Sigma}_{gl}$ and w_{gl} for $l = 1, \ldots, m_g$ and $g = 1, \ldots, G$; this is a quite large number of parameters as it is

$$h_g^{GMM} = m_g p + m_g \frac{p(p+1)}{2} + (m_g - 1), \qquad \text{for } g = 1, \ldots, G$$

which may be difficult to estimate for relatively small sample sizes.

Hastie and Tibshirani [Hastie and Tibshirani, 1996] introduce what they call mixture discriminant analysis (MDA) which exploits Gaussian mixtures for classification purposes by imposing some constraints which make estimation and interpretation easier. A completely different solution, aimed at reducing the number of free parameters in a mixture model, is due to McLachlan *et al.* [McLachlan *et al.*, 2002] who propose to assume a factor model for each mixture component, thus modeling the density as a mixture of factor analyzers.

In this paper we derive an approach for modeling class conditional densities which combines the potentialities of the independence condition in a low dimensional latent space (in the spirit of Amato *et al.*) with the semiparametric structure of mixture models. The method which simultaneously allows to address both aspects is Independent Factor Analysis [Attias, 1999].

2 Independent Factor Analysis

Independent Factor Analysis has been recently introduced by Attias (1999) in the context of signal processing and only recently it has been given a solid statistical foundation [Montanari and Viroli, 2004]. Its aim is to describe p observed variables x_j, which are generally correlated, in terms of a smaller set of k unobserved independent latent variables y_i and an additive specific term u_j:

$$x_j = \sum_{i=1}^{k} \lambda_{ji} y_i + u_j,$$

where $j = 1, ..., p$, $i = 1, ..., k$. In a more compact form the model is

$$\mathbf{x} = \Lambda \mathbf{y} + \mathbf{u} \tag{3}$$

where the factor loading matrix $\Lambda = \{\lambda_{ji}\}$ is also termed as *mixing matrix.* Its structure closely resembles the classical factor model but it differs from it as far as the properties of the latent variables it involves is concerned. The random vector $\mathbf{u}$ representing the noise is assumed to be normally distributed, $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \Psi)$ with Ψ allowing for correlations between the error terms. The latent variables $\mathbf{y}$ are assumed to be mutually independent and not necessarily normally distributed; their densities are indeed modeled as mixtures of Gaussians. The independence assumption allows to model the density of each y_i in the latent space separately. In more formal terms each factor is thus described as a mixture of m_i gaussians with mean $\mu_{i,q}$, variance $\nu_{i,q}$ and mixing proportions $w_{i,q}$ $(q = 1, ..., m_i)$:

$$f(y_i) = \sum_{q=1}^{m_i} w_{i,q} \phi \left(y_i; \mu_{i,q}, \nu_{i,q}\right) \tag{4}$$

The mixing proportions $w_{i,q}$ are constrained to be non-negative and sum to unity.

A particular characterization of the IFA model is that it involves two layers of latent variables: besides the factors, $\mathbf{y}$, an *allocation variable*, $\mathbf{z}$, must be introduced, as always when dealing with mixture models. With reference to a particular factor i, the mixture can be thought of as the density of an heterogeneous population consisting of m_i subgroups. For each observation the allocation variable denotes the identity of the subgroup from which it is drawn. In the k-dimensional space, the multivariate allocation variable, $\mathbf{z}$, follows a multivariate multinomial distribution. The density of the observed data can be constructed by conditioning to these two latent layers:

$$\begin{aligned} f\left(\mathbf{x}|\Theta\right) &= \sum_{\mathbf{z}} \int f(\mathbf{x,y,z}|\Theta) d\mathbf{y} \\ &= \sum_{\mathbf{z}} \int f(\mathbf{z}|\Theta) f(\mathbf{y}|\mathbf{z},\Theta) f(\mathbf{x}|\mathbf{y},\mathbf{z},\Theta) d\mathbf{y} \\ &= \sum_{\mathbf{z}} f(\mathbf{z}|\Theta) f(\mathbf{x}|\mathbf{z},\Theta) \qquad (5) \end{aligned}$$

where Θ denotes the whole set of the IFA model parameters.

It is not difficult to derive that the conditional density $f(\mathbf{x}|\mathbf{z},\Theta)$ follows a Gaussian distribution since it is the convolution of two Gaussian densities:
.

$$\mathbf{x}|\mathbf{y},\mathbf{z} \sim \mathcal{N}(\Lambda\mathbf{y},\Psi) \qquad (6)$$

and

$$\mathbf{y}|\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}},\mathbf{V}_{\mathbf{z}}) \qquad (7)$$

where $\boldsymbol{\mu}_{\mathbf{z}}$ and $\mathbf{V}_{\mathbf{z}}$ are respectively defined as:

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

For more details see [Montanari and Viroli, 2004].

Therefore the expression (5) indicates that the density of the observed data given the IFA model, *i.e.* the likelihood function $f(\mathbf{x}|\Theta)$, is a finite mixture of p-variate normals. Its generic component is given by

$$f(\mathbf{x}|\mathbf{z},\Theta) = \phi\left(\mathbf{x}|\mathbf{z};\Lambda\boldsymbol{\mu}_{\mathbf{z}},\Lambda\mathbf{V}_{\mathbf{z}}\Lambda^T + \Psi\right) \qquad (8)$$

Implicit in the IFA estimation problem (which is solved by the EM-algorithm) are the two assumptions regarding the correct number of factors and the number of mixture components for modeling each factor. Assessing the correct specification of the model is an important but as jet unsolved problem; this issue has been addressed in [Montanari and Viroli, 2004]. Once the number of factors, k, and the number of components for each of them, m_i, have been correctly specified, the total number of the IFA model parameters for f_g is

$$h_g^{IFA} = pk + \frac{p(p+1)}{2} + 3\sum_{i=1}^{k} m_i - k \qquad g = 1, \dots, G.$$

As a consequence of this *a priori* double choice, the number of the mixture components in (5) is univocally determined as $m_g = \prod_{i=1}^{k} m_i$.

An advantage of this approach is that by applying mixture models not directly to the observed variables but onto the reduced latent space the density of the observed variables is still a Gaussian mixture model that generally involves a smaller set of parameters. For instance, a mixture model for the class conditional density f_g with $m_g = 4$ components can be obtained by estimating $k = 2$ factors with $m_i = 2$ components each. The resulting number of parameters

$$h_g^{IFA} = \frac{p^2}{2} + \frac{5}{2}p + 10$$

is smaller than the one to be estimated in (2)

$$h_g^{GMM} = 2p^2 + 6p + 3$$

for $p \geq 2$.

A further appealing feature of the proposed solution is that the formulation given by (4) and (5) does not rely on any constraints on the parameters, allowing for a very flexible density approximation.

3 Empirical results

3.1 Simulated data

The discrimination performance of Independent Factor Discriminant Analysis (IFDA) has been tested on 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 $i = 1, \dots, 21$, 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 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)
IFDA (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 eight entries above the line are taken from Hastie and Tibshirani (1996). The last line indicates the error rates in the IFDA with 2 components for each factor.

Table 1 indicates the classification results taken from Hastie and Tibshirani [Hastie and Tibshirani, 1996] and includes the performances of IFDA over 10 simulations. Independent Factor Discriminant Analysis shows the lowest classification error rate in the test samples.

3.2 Real data

We 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 *et al.*, 1994], only one split into training and test set has been considered.

Independent Factor 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
IFDA (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 the IFDA with 2 components for each factor.

4 Conclusion

In this paper we have proposed a new approach to classification by Gaussian mixtures. Its main assumption is that the observed data have been generated by an independent factor model. In this way we obtain a very flexible density approximation, which, for a given number of mixture components, is often based on a lesser number of parameters than the classic mixture model solution and allows for heteroscedastic components. Its performance seems to be very competitive with respect to the main classification procedures proposed in the statistical literature.

References

[Amato *et al.*, 2002]U. Amato, A. Antoniadis, and Gréfoire G. Independent component discriminant analysis. *International Mathematical Journal*, pages 735–753, 2002.

[Attias, 1999]H. Attias. Independent factor analysis. *Neural Computation*, pages 803–851, 1999.

[Breiman *et al.*, 1984]L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees.* Wadsworth, Belmont, California, 1984.

[Comon, 1994]P. Comon. Independent component analysis, a new concept?. *Signal Processing*, pages 287–314, 1994.

[Coomans *et al.*, 1983]D. Coomans, M. Broeckaert, and D.L. Broeckaert. Comparison of multivariate discriminant techniques for clinical data - application to the tyroid functional state. *Meth. Inform. Med.*, pages 93–101, 1983.

[Hand, 1982]D.J. Hand. *Kernel Discriminant Analysis.* Research Studies Press, Letchworth, 1982.

[Hastie and Tibshirani, 1996]T. Hastie and R. Tibshirani. Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 155–176, 1996.

[Hastie *et al.*, 1994]T. Hastie, R. Tibshirani, and A. Buja. Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association*, pages 1255–1270, 1994.

[McLachlan and Peel, 2000]G.J. McLachlan and D. Peel. *Finite Mixture Models.* Wiley, New York, 2000.

[McLachlan *et al.*, 2002]G.J. McLachlan, R.W. Bean, and D. Peel. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, pages 413–422, 2002.

[Montanari and Viroli, 2004]A. Montanari and C. Viroli. The independent factor analysis approach to latent variable modeling. *Submitted*, 2004.

[Polzehl, 1995]J. Polzehl. Projection pursuit discriminant analysis. *Computational Statistics and Data Analysis*, pages 141–157, 1995.

[Silverman, 1986]B.W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London, 1986.