Reflection on a supervised approach to Independent Component Analysis

Cinzia Viroli

Dipartimento di Scienze statistiche
Università di Bologna, I-40126 Bologna, Italy

Abstract. This work focuses on a recent supervised approach to Independent Component Analysis, a linear transformation method that provides latent variables assumed to be non-gaussian and mutually independent. According to this approach the latent structure is identified by estimating the joint product density of independent components, using a technique that transforms the unsupervised learning problem into a supervised function approximation one. Like projection pursuit methodology, this procedure attempts to get interesting projections of the observed units, that seem to capture the latent clustered structure of the data.

1 Introduction

One of the aims of exploratory data analysis is the identification of underlying structure, in order to detect interesting representations of multivariate data for purpose of classification or clustering; such a representation is often sought as a linear transformation of the observed data.

A recent multivariate method developed for discovering linear or non linear projections of the data onto a lower dimensional space is Independent Component Analysis (ICA). In this context, the latent variables are assumed to be non-gaussian and mutually independent. In terms of exploratory data analysis, interesting non-gaussian distributional forms may be sub- or super-gaussian, bi-modal or multi-modal. In the particular context of classification we focus our attention on multi-modal latent components, since they may indicate specific clusters and classes inherent in the data.

While many popular approaches to ICA are based on suitable choice of a non-normality or dependence index, the aim of this paper is to test a different approach to ICA recently proposed. According to this procedure, the latent structure is identified by viewing the density estimation task as a two class classification problem, after that a particular trick to transform the problem into a supervised form has been applied. The built model can be interpreted as a generalized projection pursuit regression. The approach is tested to simple classification problems.

2 Independent Component Analysis

Denote by $x_1, x_2, ...x_p$ the observed variables, which are supposed to be modelled as linear combinations of q hidden variables $y_1, y_2, ..., y_q$:

$$x_i = a_{i1}y_1 + a_{i2}y_2 + ... + a_{iq}y_q \qquad \text{for all} \quad i = 1, ..., p \tag{1}$$

where the a_{ij} $(j = 1, ..., q)$ are some real coefficients. By definition, the y_i are *statistically mutually independent*. The basic ICA model can put in the following compact formulation:

$$\mathbf{X} = \mathbf{AY}. \tag{2}$$

Since it describes how the observed data are generated by a mixing process of hidden components, the matrix $\mathbf{A}$ is often called *mixing matrix*. In the standard ICA model we have $p = q$ and thus the matrix $\mathbf{A}$ is square. In the model both the mixing matrix and latent components are unknown and can be estimated under some precise restrictions. The fundamental assumption on which ICA rests is the independence of the latent components. As clearly shown in Hyvärinen *et al.* (2001), this condition implies that the research of a distributional form for the hidden projections moves onto the direction of non-gaussianity. In particular, the model is identifiable when at most one of the q independent components is gaussian.

The current algorithms for estimating independent components can be divided into procedures based on minimization or maximization of some relevant criterion functions (mutual information, kurtosis, negentropy...) and methods based on the stochastic gradient scheme which may have implementation in neural networks. Further details on these algorithms and their convergence property can be found in Hyvärinen *et al.* (2001) and Hyvärinen and Oja (1997). All these procedures have the form of *unsupervised learning*, since they do not involve the use of some outcome variables to guide the learning process. An interesting alternative is represented by the introduction of a new dichotomic variable that allows to revaluate the unsupervised learning problem as a *supervised* function approximation one (Hastie *et al.*, 2001). The starting point is to consider the ICA model as the problem of marginalizing the joint density of the observed variables. If the y_i are *statistically mutually independent* their joint density function is factorizable:

$$h(\mathbf{y}) = \prod_{i=1}^{p} h_i(y_i) \tag{3}$$

According to the probabilistic result on the density of a transformation, the basic ICA model can be formulated in terms of density function estimation:

$$g(\mathbf{x}) = \left|\det\left(\mathbf{A}^{-1}\right)\right| h(\mathbf{y}) = \left|\frac{1}{\det\mathbf{A}}\right| \prod_{i=1}^{p} h_i(y_i) = \prod_{i=1}^{p} h_i(y_i) \tag{4}$$

where the last equality is true when the mixing matrix $\mathbf{A}$ is an orthogonal matrix.

3 Unsupervised as supervised learning

Let $\mathbf{X}$ be the multivariate set of observed data, with joint probability density function $g(\mathbf{x})$. From equation (4) $g(\mathbf{x})$ has to be factorizable as the product of p unknown density functions. Let $g_0(\mathbf{x})$ be a specific known density function, used for reference. Suppose that the observed sample $\mathbf{X}$ has size N; then another sample of the same size N is drawn from $g_0(\mathbf{x})$ using Monte Carlo methods. The idea is to resample the same observed pattern $\mathbf{x}_n$ with $n = 1, ..., N$ using the (different) probability of the reference density function $g_0(\mathbf{x})$. The two data sets $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ and $\{\mathbf{x}_{N+1}, \mathbf{x}_{N+2}, ..., \mathbf{x}_{2N}\}$ are *i.i.d.* random samples respectively from $g(\mathbf{x})$ and $g_0(\mathbf{x})$. These samples are pooled and a new random variable Y is created, assigning $Y = 1$ to those observations drawn from $g(\mathbf{x})$ and $Y = 0$ to those drawn from the reference density function $g_0(\mathbf{x})$. The conditional expectation of Y given $\mathbf{x}$ is:

$$\mu(\mathbf{x}) = E(Y|\mathbf{x}) = \frac{g(\mathbf{x})}{g(\mathbf{x}) + g_0(\mathbf{x})}. \tag{5}$$

This quantity can be approximate using the combined sample $(y_1, \mathbf{x}_1)$, $(y_2, \mathbf{x}_2)$ ..., $(y_{2N}, \mathbf{x}_{2N})$, where the variable Y can be reinterpreted as outcome variable. Hence, the posterior mean $E(Y|\mathbf{x})$ may be estimated by a supervised learning.

Moreover, the identity (5) can be solved for $g(\mathbf{x})$:

$$\hat{g}(\mathbf{x}) = g_0(\mathbf{x}) \frac{\hat{\mu}(\mathbf{x})}{1 - \hat{\mu}(\mathbf{x})}. \tag{6}$$

Applying the logarithm to the previous expression, the logit of $\mu(\mathbf{x})$ can be viewed as the difference:

$$\text{logit}\,(\mu(\mathbf{x})) = \log g(\mathbf{x}) - \log g_0(\mathbf{x}). \tag{7}$$

Thus, $\mu(\mathbf{x})$ provides information concerning departures of the data density $g(\mathbf{x})$ from the chosen reference density $g_0(\mathbf{x})$ and the $\text{logit}\,(\mu(\mathbf{x}))$ can be viewed as a "contrast" statistic. This expression offers a suggestion for choosing the form of the reference density function: a good choice for $g_0(\mathbf{x})$ is dictated by types of departures that are most interesting. In the particular context of ICA the aim is the estimation of independent components and hence the *departure from dependence* is investigated. Since independent components means non-gaussian components, this is equivalent investigating *departure from joint normality*. For this purpose a good choice for $g_0(\mathbf{x})$ could be the multivariate gaussian distribution.

Following the result in the equation (4), the aim is to factorize the expression (6) as a product:

$$\hat{g}(\mathbf{x}) = g_0(\mathbf{x}) \frac{\hat{\mu}(\mathbf{x})}{1 - \hat{\mu}(\mathbf{x})} = h_1(a_1^T \mathbf{x}) h_2(a_2^T \mathbf{x}) ... h_p(a_p^T \mathbf{x}) \tag{8}$$

where the components $a_1^T\mathbf{x}, a_2^T\mathbf{x}, ..., a_p^T\mathbf{x}$ are independent.

In order to factorize the reference function a change of variable is needed and it is easy to see that $g_0(a_1^T\mathbf{x}, ..., a_p^T\mathbf{x})$ is factorizable if $\{a_1, a_2, ..., a_p\}$ is a set of orthogonal vectors and $g_0(\mathbf{x})$ is a spherical multivariate gaussian distribution. This choice confirms all the previous implications of the (7). Secondly, in order to estimate the logit $(\mu(\mathbf{x}))$, Hastie *et al.* (2001) proposed an additive logistic regression:

$$\text{logit}\,(\mu(\mathbf{x})) = f_1(a_1^T\mathbf{x}) + f_2(a_2^T\mathbf{x}) + ... + f_p(a_p^T\mathbf{x}). \tag{9}$$

In this case:

$$\frac{\mu(\mathbf{x})}{1-\mu(\mathbf{x})} = \exp\left\{f_1(a_1^T\mathbf{x})\right\} \cdot \exp\left\{f_2(a_2^T\mathbf{x})\right\} ... \cdot \exp\left\{f_p(a_p^T\mathbf{x})\right\}. \tag{10}$$

As the authors declared "*... while this procedure appears to work well on some simple examples, at the time of writing it is largely untested*". In this work the logit $(\mu(\mathbf{x}))$ is modelled by a generalized projection pursuit regression. This choice is dictated from the intention to emphasize the analogies with the exploratory projection pursuit. Moreover, the generalized projection pursuit regression is more general and flexible than the additive logistic regression, as explained in the following section.

4 Generalized Projection Pursuit Regression

Friedman, J.H. and Stuetzle, W. (1981) defined Projection Pursuit Regression as a model in which the response is related to a sum of smooth functions of linear combinations of the covariates:

$$E[Y|X_1, X_2, ..., X_p] = \beta_0 + \sum_{j=1}^{q} \beta_j f_j(\alpha_j^T\mathbf{x}). \tag{11}$$

In order to be estimable, the model has to satisfy normalization and standardization conditions such as:

$$\sum_{k=1}^{p} \alpha_{kj}^2 = 1, \quad E(f_j) = 0, \quad E(f_j^2) = 1 \qquad \forall j = 1, ..., q.$$

We can generalize the projection pursuit model by allowing two extensions: the first one is that the distributional form of the response Y may come from an exponential family and secondly the relation between the response variable Y and its predictor is not necessarily the identity link but may be any monotonic differentiable function. Therefore Generalized Projection Pursuit Regression (GPPR) consists of three components:

- random component; the functional form of the response variable Y may be
$$f_Y(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{\alpha(\phi)} + c(y,\phi)\right\}$$
where θ is the natural parameter and ϕ the scale parameter.
- systematic component; a weighted sum of smooth functions of linear combinations of covariates produce a predictor of Y
$$\eta(X) = \beta_0 + \sum_{j=1}^{q} \beta_j f_j(\alpha_j^T \mathbf{x})$$
- link component; the relation between the predictor $\eta(X)$ and the expected value μ of the response Y may be any monotonic differentiable function $g(E[Y|X]) = \eta(X)$.

5 Logistic Projection Pursuit Regression

As a particular case of the Generalized Projection Pursuit Model, Logistic Projection Pursuit Regression models a binary response. In LPP model the response Y has a bernoulli distribution of parameter μ.

Since we have $0 < \mu < 1$ the link should satisfy the condition that it maps the interval $(0, 1)$ onto the whole real line. The link component is the logistic function:

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \sum_{j=1}^{q} \beta_j f_j(\alpha_j^T \mathbf{x}) = \eta(X).$$

Parameter and function estimation is obtained by using a modified version of a local scoring algorithm that combines the iteratively reweighted least squares (IRLS) with a back-fitting procedure. Schematically the IRLS algorithm consists of looping until convergence the following two steps:

- calculate an adjusted dependent variable z with estimated variance w^{-1};
- regress z on the predictors with weights w.

The adjusted dependent variable is obtained by approximating $g(y)$ with a first-order Taylor series expansion about μ. In the logistic regression problem the adjusted responses z are:

$$z = \eta + \frac{y-\mu}{\mu(1-\mu)} \qquad \text{with weights} \quad w = \mu(1-\mu).$$

The algorithm used to estimate parameters and functions in the logistic projection pursuit regression is a modified version of Roosen and Hastie proposal (1993) and is based on the following steps:

- Initialize variables.
- Estimate all terms.
- Prune terms of least importance.
- Clean up.

In the initialization step, all the coefficients α_j are chosen randomly from the uniform distribution in $[-1, 1]$ and normalized. The smooth functions are set identically equal to zero. The parameters β_j are fixed equal to one. The initial choice of parameters and functions does not seem to have great influence on the convergence time.

In the second step, a total of $p \geq q$ terms is fitted and a backward selection procedure is then used to prune down to q terms. For each term the adjusted dependent variable z with weights w is calculated. Then the partial residuals r_i are obtained as difference between z_i and previous terms (with $i = 1, ..., 2N$). Finally each term is updated with response r_i and weights w_i. The procedure is realized under the condition that the estimated vectors $\hat{\alpha}_j$ (with $j = 1, ..., p$) are orthonormal. This implies that it may be necessary to orthogonalize the resulting mixing matrix after each iteration, projecting the current estimated vector $\hat{\alpha}_j$ onto the orthogonal space generated from the previous estimated set $\{\hat{\alpha}_1, ..., \hat{\alpha}_{j-1}, \hat{\alpha}_{j+1}, ..., \hat{\alpha}_p\}$.

In the prune step, terms of least importance are dropped and then only q terms are updated using the backfitting procedure within local scoring. In LPP estimation we can consider the terms $\alpha_j^T \mathbf{x}$ as estimate of each independent component. To measure the importance of each term we can use the magnitude of the scale coefficient β_j or we can base on its t-statistics from the regression of y on $\{f_m(\nu_m)\}_{m=1}^j$. A further possibility is using the *negentropy*, a relevant concept of information theory. Let X be a random variables with probability density measured by $p(X)$. Let $p(X_G)$ the gaussian distribution with the same mean and variance-covariance matrix as $p(X)$. The negentropy is a particular relative entropy:

$$J(X) = \int p(x) \log \frac{p(x)}{p(x_G)} dx = H(X_G) - H(X), \tag{12}$$

where $H(X)$ and $H(X_G)$ are the entropies of X and X_G respectively. The idea of using negentropy in this context is based on the following property: a gaussian variable has the largest entropy among all the random variables of equal variance. Hence, negentropy is always non-negative and it is zero if X has a gaussian distribution. This means that we can use negentropy to measure the departure from the gaussianity. In practice this is obtained by dropping sequentially the estimated component with the lowest negentropy.

Finally, the clean up procedure consists of adjusting each estimated smooth function to have zero mean, changing the value of coefficient β_0 accordingly.

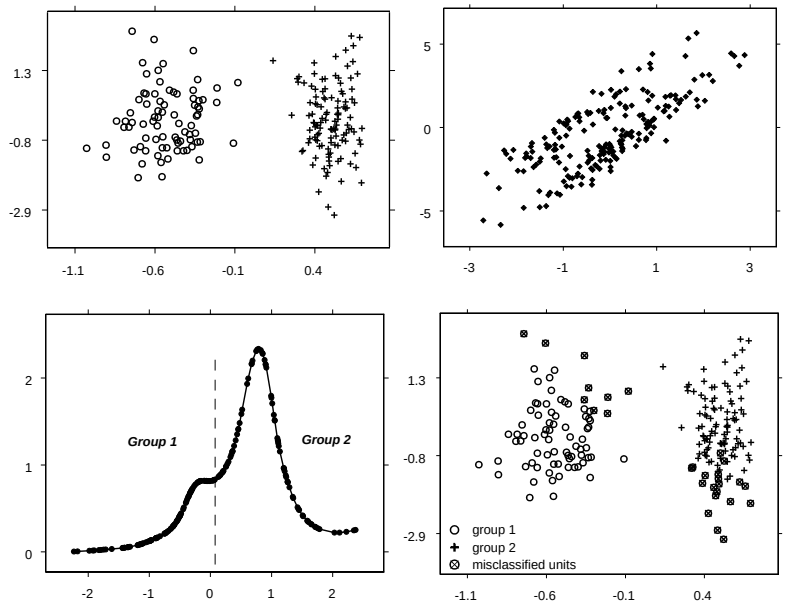

Fig. 1. A simple simulation.

6 Some applications to classification problems

The procedure has been applied to a simple simulation and to the Iris data set. The first example involves two latent variables, assumed to be a mixture of gaussians and a standard gaussian respectively:

$$f(y_1) = 0.4\mathcal{N}(-0.5, 0.2) + 0.6\mathcal{N}(0.5, 0.1) \quad \text{and} \quad f(y_2) = \mathcal{N}(0, 1).$$

The two independent variables are linearly mixed as follows:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

The first diagram in Figure 1 shows the joint distribution of the two latent variables. It's clear from this graph that there are two distinct groups. The second graph shows the joint distribution of the observed mixed variables, after some noise is added. In this second graph the discrimination between the two groups is not so clear. Note that, in this particular example, the application of principal component analysis does not allow to distinguish the two groups, since the first component lies in the direction of maximum variance and captures almost all the variability. The estimation by ICA gives two latent components but the second one is dropped because it is too much close to the normal distribution. The third graph shows the density distribution of the only one estimated component, that is clearly bimodal. It is possible to

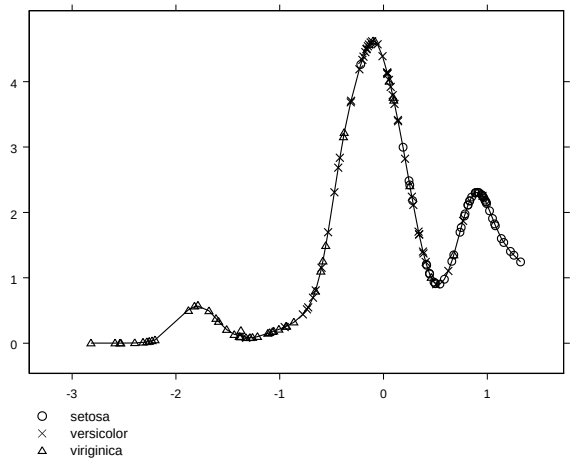

Fig. 2. ICA on IRIS dataset.

discriminate between the two groups by considering the minimum between the two modes. The last graph presents the estimated classification using the clustered latent structure: 28 units out of 200 are misclassified.

The application on the Iris dataset is also interesting. The unique non-gaussian component presents three modes, as shown in Figure 2. In this case 26 units out of 150 are misclassified. The largest misclassification corresponds to the distinction between the species *versicolor* and *virginica*: this is a usual problem encountered using this dataset in the context of the linear discrimination.

7 Conclusions

This supervised approach to ICA allows one to emphasize its relation with the exploratory projection pursuit. In particular, under the constraint of orthogonality of the mixing matrix, the previous procedure shows that independent component analysis is substantially equivalent to the aim of searching for interesting (and so non-gaussian) projection pursuit directions. In this spirit, the system of orthogonal coefficients of the mixing matrix can be properly reinterpreted as system of projection indexes.

The procedure encounters some problems when the dimension of the observed variables is high. This is due to the fact that the observed data have to be resampled from a multivariate gaussian distribution by Monte Carlo methods. To better understand the nature of the problem suppose a simple situation in which the observed data are generated from a uniform distribution in $[-3, 3]$ of dimensionality from 1 to 5 respectively. The marginal uniform random variables are supposed to be mutually independent and therefore the

covariance matrix is diagonal. The simulated points are then resampled following the probabilities of a multivariate normal standard distribution. We compute the theoretical frequencies of the central interval $[\mu - \sigma, \mu + \sigma]$ in both distributions. The results are presented in the table (1).

Dimension	*Normal*	*Uniform*
1	0.683	0.577
2	0.466	0.333
3	0.318	0.192
4	0.217	0.111
5	0.148	0.064

Table 1. *Areas of the central interval* $[\mu-\sigma, \mu+\sigma]$ *in the uniform and the standard normal distribution with different dimensionality.*

In this particular situation we see that with only 5 observed variables the probability of having an observed value near the mean is only equal to 0.064. In the simulation procedure the rare points in the central interval must be repeated lots of times in order to reflect the the higher theoretical probabilities of the multivariate normal distribution. The risk is not to observe any value "near" the mean, so that all simulated points lie onto the tails of the multivariate normal distribution. Therefore we don't evaluate the departures from the entire body of the normal distribution but just from its tails. The seriouness of the problem depends on the number of the observed variables and specially on the *real* distribution from which the observed data are generated.

References

FRIEDMAN, J.H. and STUETZLE, W. (1981): Projection Pursuit Regression. *Journal of the American Statistical Association, 76, 817-823.*

GIROLAMI, M. CICHOCKI, A. and AMARI, S. (1998): A common Neural Network Model for Unsupervised Exploratory Data Analysis and Independent Component Analysis. *IEEE Transaction on Neural Netowoks.*

HASTIE, T. and TIBSHIRANI, R. (1990): *Generalized Additive Models.* Chapman and Hall.

HASTIE, T. TIBSHIRANI, R. and FRIEDMAN, J.H. (2001): *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer.

HYVÄRINEN, A. KARHUNEN, J. and OJA, E. (2001): *Independent Component Analysis.* John Wiley & Sons, INC.

HYVÄRINEN, A. and OJA, E. (1997): A Fast Fixed-Point Algorithm for Independent Component Analysis. *Neural Computation, 9, 1483-1492.*

ROOSEN, C.B. and HASTIE, T. (1993): Logistic Response Projection Pursuit. *AT&T Bell Labs Technical Report.*