Independent Component Analysis for Non-Normal Factor Analysis

Aapo Hyvärinen[1] and Yutaka Kano[2]

[1] Neural Networks Research Centre, Helsinki University of Technology, Finland
[2] School of Human Sciences, Osaka University, Japan

Summary. Independent component analysis (ICA) was developed in the signal processing and neural computation communities. Its original purpose was to solve what is called the blind source separation problem: when linear mixtures of some original source signals are observed, the goal is to recover the source signals, using minimum assumptions on the mixing matrix (i.e. blindly). This leads to a linear model that is very similar to the one used in *factor analysis*. What makes ICA fundamentally different from conventional factor analysis is that the source signals are assumed to be *non-Gaussian*, in addition to the basic assumption of their independence. In fact, this implies that the model can be uniquely estimated from the data, using supplementary information that is not contained in the covariance matrix. Interestingly, a very close connection can be found with *projection pursuit*: The basic form of the ICA model can be estimated by finding the projections that are maximally non-Gaussian, which is the goal of projection pursuit as well. On the other hand, the dimension of the observed data vector is often first reduced by principal component analysis, in which case ICA can be viewed as a method of determining the *factor rotation* using the non-Gaussianity of the factors.

Keywords: Factor analysis, independent component analysis, projection pursuit, factor rotation, non-normality

1 Introduction

Independent Component analysis (ICA) is a multivariate linear latent variable model. In its formulation, it is very closely related to factor analysis (see e.g., Lawley and Maxwell, 1971) which was developed mainly by social scientists. Its actual estimation methods, on the other hand, are very similar to projection pursuit, developed by statisticians (see e.g., Huber, 1985). The key difference between ICA and ordinary factor analysis is that the latent factors are assumed to be *non-Gaussian*, i.e. to have non-normal distributions. This seemingly small difference in the model definition leads to huge differences in the estimation procedure and the applications of the model. In fact, non-normality allows us to *separate* several linearly mixed independent latent *signals*, and also to *uniquely determine the factor rotation* without traditional factor rotation methods such as varimax. The purpose of this paper is to introduce ICA to a reader that is already familiar with factor analysis as it is usually applied in the social sciences.

The history of ICA goes back to the early 80's when Hérault, Jutten and Ans (Hérault and Ans, 1984; Jutten and Hérault, 1991) considered a problem in computational neuroscience: How is it possible that when neural fibres carry signals that are mixtures of some underlying source signals, the central nervous system is able to recover (separate) those source signals. A small group of researchers, mainly French, developed the basic idea further in a signal processing context. Possibly the first principled estimation method for ICA was proposed by Cardoso (1989), and Comon (1994) laid the theoretical foundation in his fundamental paper showing that the model was identifiable in the sense that everybody had hoped for. After 1995, ICA was enthusiastically received by people working on neural networks and computational neuroscience due to the work by Bell and Sejnowski (1995), who developed an improved algorithm for ICA estimation, and Olshausen and Field (1996), who showed explicit connections between ICA and models of the visual processing in the brain.

The basic definition of ICA is surprisingly simple. Let $x_1, x_2, ..., x_n$ denote n observed random variables. These are modelled as a linear transformation of n latent variables $s_1, s_2, ..., s_n$:

$$x_i = \sum_{j=1}^{n} a_{ij} s_j, \text{ for } i = 1, 2, ..., n. \tag{1}$$

The a_{ij} are constant unknown parameters to be estimated, not unlike factor loadings. We make the following assumptions on the latent variables or independent components s_i:

1. The s_i are mutually (statistically) independent.
2. The s_i are non-Gaussian, i.e. have non-normal distributions.

The linear mixing model in Eq. (1) is not terribly different from an ordinary factor analysis model. One difference is that there are no separate noise variables or specific factors in this model. However, the number of the factors is quite large, in fact, equal to the number of observed variables. Thus, we could think that the common and specific factors, as well as noise are just all grouped together in the s_i.[1]

Assumption 1 is not unlike the one usually made in factor analysis in the case of maximum likelihood estimation. If the s_i follow a joint Gaussian distribution, independence follows from the conventional assumption of uncorrelatedness. This is a special property of the normal distribution, however, and we shall see below that in the case of non-Gaussian variables, uncorrelatedness does *not* at all imply independence.

[1] Denoting the number of common factors be k, the total number of latent factors including specific ones becomes $n + k$ in ordinary factor analysis. This is always larger than the maximum number of independent components in ICA, which is n. However, if n is large and k is small, the difference may not be very important.

So, what really distinguishes the ICA model from ordinary factor analysis is the second assumption of non-Gaussianity. In fact, a possibly more illuminating name for ICA would be *non-Gaussian factor analysis.* Due to non-Gaussianity, both the estimation theory and practical results of ICA are very different from those obtained by ordinary factor analysis.

In the next section, we will first show how ICA is able to do "blind source separation", something that ordinary factor analys is not able to do. In section 3 we will discuss why this is so, why non-Gaussianity is so important, and how ICA can be interpreted as a factor rotation. Section 4 discusses basic statistical criteria that can be used to estimate the ICA model, and shows the intimate connection between ICA and projection pursuit. Finally, Section 5 concludes the paper.

2 Blind Source Separation

To see the drastic effect of the assumption of non-Gaussianity, let us consider a problem that has inspired a large part of ICA research, called blind source separation. Imagine that you are in a room where a number of people (say, three) are speaking simultaneously. You also have three microphones, which you hold in different locations. The microphones give you three recorded time signals, which we could denote by $x_1(t), x_2(t)$ and $x_3(t)$, with x_1, x_2 and x_3 the amplitudes, and t the time index. Each of these recorded signals is a weighted sum of the speech signals ("sources") emitted by the three speakers, which we denote by $s_1(t)$, $s_2(t)$, and $s_3(t)$. We could express this as a linear equation which is just like Eq. (1), where the a_{ij} with $i, j = 1, ..., 3$ are some parameters that depend on the distances of the microphones from the speakers. The goal in blind source separation is to estimate the original speech signals $s_1(t), s_2(t)$, and $s_3(t)$, using only the recorded signals $x_i(t)$.

As an illustration, consider the waveforms in Fig. 1. The original speech signals could look something like those on the left, and the mixed signals could look like those in the middle. The problem is to recover the "source" signals using only the mixed data.

One approach to solving this problem would be to use some information on the statistical properties of the signals $s_i(t)$ to estimate both the a_{ij} and the $s_i(t)$. Let us assume that $s_1(t), s_2(t)$, and $s_3(t)$ are, at each time instant t, statistically independent. If the s_i are non-Gaussian as well, we see that we have in fact all the assumptions of the ICA model! Independent component analysis can thus be used to estimate the a_{ij}, and this allows us to separate the three original signals from their mixtures. This is called "blind" source separation because hardly any information on the sources was used, only the very weak assumptions on statistical independence and non-Gaussianity.

Figure 1, on the right, gives the three signals estimated by ICA. As can be seen, these are very close to the original source signals on the left (the signs of some of the signals are reversed, but this has no significance.)

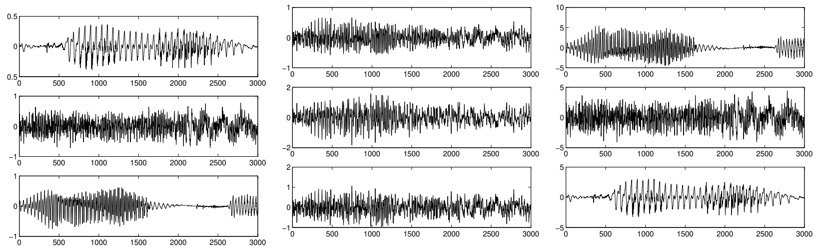

Fig. 1. Left: The original audio signals. Middle: The observed mixtures of the original signals. Right: The estimates of the original signals, obtained by ICA.

3 ICA vs. Factor Analysis

Factor analysis does not separate sources It is important to note that ordinary factor analysis cannot separate source signals as described in the preceding section. This is because factor analysis, or related techniques such as principal component analysis, can only estimate the factors up to a rotation. But in the preceding source separation example, we had three source signals, that is, three factors, and also three observed variables. If one is able to estimate the factors (source signals) only up to a rotation, that means that one is not really able to estimate them at all, since most orthogonal rotations mix the source signals just as badly as the original mixing.

It may be very surprising that the original sources or independent components can be recovered at all. Indeed, the proof that this is possible was presented only relatively recently (Comon, 1994), and it certainly does surprise many people hearing it for the first time. In the following, we will try to explain intuitively why the non-Gaussianity assumption enables the estimation of the model.

Illustration of why ICA is possible To illustrate the ICA model in statistical terms, consider two independent components that have uniform distributions The joint density of s_1 and s_2 is then uniform on a square, which is illustrated in Fig. 2, on the left.

Now, assume that the mixing matrix $\mathbf{A}$ is orthogonal. Basically, we make this assumption here because we consider the problem of estimating an orthogonal factor rotation (see below). Mixing these variables, we obtain the observed data $\mathbf{x}$ as shown in Fig. 2, in the middle. The mixed data has a uniform distribution on a rotated square. Actually, from Fig. 2 you can see an intuitive way of estimating $\mathbf{A}$: The *edges* of the square are in the directions of the columns of $\mathbf{A}$. This means that we could, in principle, estimate the ICA model by first estimating the joint density of x_1 and x_2, and then locating the edges. So, intuitively we see that the problem can be solved, in this special case.

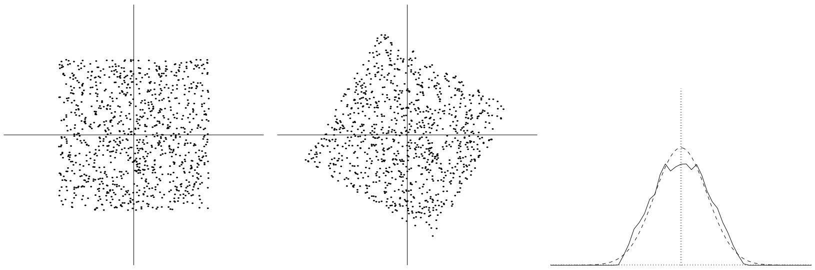

Fig. 2. Left: The joint distribution of the independent components s_1 and s_2 with uniform distributions. Middle: The joint distribution of the observed (orthogonal) mixtures x_1 and x_2. Right: The marginal distribution of a mixture.

Why Gaussian variables are no good On the other hand, we can illustrate why Gaussian variables are not allowed. Assume that the joint distribution of s_1 and s_2 is Gaussian. Using the classic formula of transforming densities we get the joint density of the mixtures x_1 and x_2 as

$$p(x_1, x_2) = \frac{1}{2\pi} \exp(-\frac{\|\mathbf{A}^T\mathbf{x}\|^2}{2})|\det \mathbf{A}^T| = \frac{1}{2\pi} \exp(-\frac{\|\mathbf{x}\|^2}{2}) \quad (2)$$

where the latter equality comes from the orthogonality of $\mathbf{A}$ (again, we consider an orthogonal factor rotation). Thus we see that the orthogonal mixing matrix does not change the density, since it does not appear in this equation at all. This means that there is no information in the observations of x_1 and x_2 that could be used to estimate $\mathbf{A}$.

This phenomenon is related to the property that uncorrelated jointly Gaussian variables are necessarily independent. Thus, the information on the independence of the components does not get us any further than uncorrelatedness. Thus, in the case of Gaussian independent components, we can only estimate the ICA model up to an orthogonal transformation, which is in fact what ordinary factor analysis does.

ICA as factor rotation In classic factor analysis, the fact that Gaussian variables leave the orthogonal transformation undetermined is well known. Many methods have been developed to determine the "factor rotation", i.e. to find a suitable orthogonal transformation. However, none of the conventional methods try to determine the rotation so that the the blind source separation problem would be solved, that is, so that the matrix $\mathbf{A}$ of factor loadings would be properly estimated. The only exception seems to be the work by Mooijaart (1985), who employed the estimation of generalized least squares using third-order moments in addition to the second-order moments to propose a new estimation procedure in factor analysis for non-normal distributions.

Conventional factor rotations use criteria that are very different from nonnormality, typically related to an easy interpretation of the factor structure.

Often in social sciences, the investigators expect there to be a relatively small number of latent factors, each of which has its indicators. That is, the indicator variables of a given factor are largely loaded on that factor, but almost unrelated with the other factors. Mathematically speaking, each row of the factor loadings matrix $\mathbf{A}$ has only one salient loading. Many rotation methods have been proposed to achieve this. However, there seems no theoretical background that justify the use of such methods.

ICA has no rotation problem, since the matrix $\mathbf{A}$ can be estimated almost completely, up to trivial scale indeterminacies. In fact, in many cases, before application of ICA, the dimension of the data is first reduced by PCA or factor analysis. ICA then gives an orthogonal rotation of the factors. Especially in that case, ICA can be seen as a factor rotation which is determined by the search for the *true factors* that really are independent.

4 Principles of ICA Estimation

"Non-Gaussian is independent"

Intuitively speaking, the key to estimating the ICA model is non-Gaussianity. The starting point here the Central Limit Theorem that says that the distribution of a sum of independent random variables tends toward a Gaussian distribution, under certain conditions. Thus, a sum of two independent random variables usually has a distribution that is closer to Gaussian than any of the two original random variables. This is illustrated in Fig. 2, on the right, where the density of a mixture is shown: it is closer to the Gaussian distribution than the uniform density is.

To estimate one of the independent components, we consider a linear combination of the x_i; let us denote this by $y = \mathbf{w}^T\mathbf{x} = \sum_i w_i x_i$, where $\mathbf{w}$ is a vector to be determined. Let us make a change of variables, defining $\mathbf{z} = \mathbf{A}^T\mathbf{w}$. Thus, y is a linear combination of the factors s_i according to Eq. (1), with weights given by z_i. Since a sum of even two independent random variables is typically more Gaussian than the original variables, $\mathbf{z}^T\mathbf{s}$ is more Gaussian than any of the s_i and becomes least Gaussian when it in fact equals one of the s_i. Therefore, we could take as $\mathbf{w}$ a vector that locally *maximizes the non-Gaussianity* of $\mathbf{w}^T\mathbf{x}$! Such a vector would necessarily correspond to a $\mathbf{z}$ which has only one nonzero component, which means that $\mathbf{w}^T\mathbf{x} = \mathbf{z}^T\mathbf{s}$ equals one of the independent components. Our approach here is rather heuristic, but it can be shown rigorously that this is a valid approach (Delfosse and Loubaton, 1995; Hyvärinen et al., 2001)

To estimate several independent components, we need to maximize the non-Gaussianities of several projections defined by vectors $\mathbf{w}_1, ..., \mathbf{w}_n$. To prevent different vectors from converging to the same maxima, it is enough to *decorrelate* the outputs $\mathbf{w}_1^T\mathbf{x}, ..., \mathbf{w}_n^T\mathbf{x}$. That is, the optimization is done under a constraint of uncorrelatedness of the $\mathbf{w}_i^T\mathbf{x}$. In the case of a factor rotation, this simply means that the rotation is orthogonal.

Our approach to ICA makes explicit the connection between ICA and *projection pursuit*. In basic projection pursuit, we try to find directions such that the projections of the data in those directions have interesting distributions, i.e., display some structure. It has been argued that the Gaussian distribution is the least interesting one, and that the most interesting directions are those that show the least Gaussian distribution. This is exactly what we do to estimate the ICA model. This also gives an interesting interpretation of what ICA is doing when the data was not generated as a sum of independent variables. Conversely, ICA gives a very illuminating characterization of projection pursuit. Typically, projection pursuit has been reported to find latent clusters and nonlinear relations, but the independence property has not been discussed at all.

Measures of non-Gaussianity

To use non-Gaussianity in ICA estimation, we must have a quantitative measure of non-Gaussianity of a random variable, say y, assumed here centred.

A classical measure of non-Gaussianity is *kurtosis*, defined as

$$\mathrm{kurt}(y) = E\{y^4\} - 3(E\{y^2\})^2. \tag{3}$$

Kurtosis is basically a normalized version of the fourth moment $E\{y^4\}$. For a Gaussian y, kurtosis is zero, and for most (but not quite all) non-Gaussian random variables, kurtosis is non-zero. Kurtosis, or rather its absolute value, has been widely used as a measure of non-Gaussianity in ICA and related fields. The main reason is its simplicity, both computational and theoretical.

However, in practice an important problem with kurtosis is that it can be very sensitive to outliers (Huber, 1985). Hyvärinen (1999) proposed a class of *robust non-Gaussianity measures*, inspired by an information-theoretic measure of non-Gaussianity, called negentropy. These measures J take the form

$$J(y) \propto [E\{G(y)\} - E\{G(\nu)\}]^2 \tag{4}$$

for some non-quadratic function G. In particular, choosing G that does not grow too fast, one obtains more robust estimators. For example, one could take $G_1(u) = \log\cosh u$, which is basically a smoother version of the absolute value, not unlike the Huber function (Huber, 1985). A very fast algorithm, called FastICA, for actually performing the optimization was proposed by Hyvärinen (1999).

Another popular approach for estimating the ICA model is *maximum likelihood estimation*. Interestingly, one can show that principles of maximum non-Gaussianity and maximum likelihood estimation are very closely connected. If the nonquadratic function G in Eq.(4) is chosen as the logarithm of the density function of the independent components (separately for each component if they have different distributions) the methods are approximately equivalent (Hyvärinen et al., 2001). For non-Gaussian components, the log-density is nonquadratic, so we see again that it is important to use information not contained in the covariance matrix.

5 Conclusion

ICA is a recently developed method for decomposing multivariate data into independent factors. The emphasis is on finding a factor rotation that gives factors that are as independent as possible. In the general case where the factors have non-normal distributions, the covariance matrix contains only a part of the information on independence, and independence is a much stronger property than mere uncorrelatedness. It can be shown that ICA is closely related to projection pursuit: the most independent factor rotation can be found by finding the most non-Gaussian projections.

References

Bell, A. and Sejnowski, T. (1995). An information-maximization approach to blind separation and blind deconvolution. Neural Computation, 7:1129–1159.

Cardoso, J.-F. (1989). Source separation using higher order moments. In Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'89), pages 2109–2112, Glasgow, UK.

Comon, P. (1994). Independent component analysis—a new concept? Signal Processing, 36:287–314.

Delfosse, N. and Loubaton, P. (1995). Adaptive blind separation of independent sources: a deflation approach. Signal Processing, 45:59–83.

Hérault, J. and Ans, B. (1984). Circuits neuronaux à synapses modifiables: décodage de messages composites par apprentissage non supervisé. Comptes.-Rendus de l'Académie des Sciences, 299(III-13):525–528.

Huber, P. (1985). Projection pursuit. The Annals of Statistics, 13(2):435–475.

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans. on Neural Networks, 10(3):626–634.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). Independent Component Analysis. Wiley Interscience.

Jutten, C. and Hérault, J. (1991). Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. Signal Processing, 24:1–10.

Lawley, D. and Maxwell, A. (1971). Factor Analysis as a Statistical Method. London: Butterworths.

Mooijaart, A. (1985). Factor analysis for non-normal variables. Psychometrica, 50:323–342.

Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature, 381:607–609.