1 Introduction

Independent component analysis (ICA) is a method for finding underlying factors or components from multivariate (multidimensional) statistical data. What distinguishes ICA from other methods is that it looks for components that are both *statistically independent*, and *nongaussian*. Here we briefly introduce the basic concepts, applications, and estimation principles of ICA.

1.1 LINEAR REPRESENTATION OF MULTIVARIATE DATA

1.1.1 The general statistical setting

A long-standing problem in statistics and related areas is how to find a suitable representation of multivariate data. Representation here means that we somehow transform the data so that its essential structure is made more visible or accessible.

In neural computation, this fundamental problem belongs to the area of unsupervised learning, since the representation must be learned from the data itself without any external input from a supervising "teacher". A good representation is also a central goal of many techniques in data mining and exploratory data analysis. In signal processing, the same problem can be found in feature extraction, and also in the source separation problem that will be considered below.

Let us assume that the data consists of a number of variables that we have observed together. Let us denote the number of variables by m and the number of observations by T. We can then denote the data by $x_i(t)$ where the indices take the values $i = 1, ..., m$ and $t = 1, ..., T$. The dimensions m and T can be very large.

A very general formulation of the problem can be stated as follows: What could be a function from an m-dimensional space to an n-dimensional space such that the transformed variables give information on the data that is otherwise hidden in the large data set. That is, the transformed variables should be the underlying *factors* or *components* that describe the essential structure of the data. It is hoped that these components correspond to some physical causes that were involved in the process that generated the data in the first place.

In most cases, we consider linear functions only, because then the interpretation of the representation is simpler, and so is its computation. Thus, every component, say y_i, is expressed as a linear combination of the observed variables:

$$y_i(t) = \sum_j w_{ij} x_j(t), \text{ for } i = 1, ..., n, j = 1, ..., m \tag{1.1}$$

where the w_{ij} are some coefficients that define the representation. The problem can then be rephrased as the problem of determining the coefficients w_{ij}. Using linear algebra, we can express the linear transformation in Eq. (1.1) as a matrix multiplication. Collecting the coefficients w_{ij} in a matrix $\mathbf{W}$, the equation becomes

$$\begin{pmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_n(t) \end{pmatrix} = \mathbf{W} \begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_m(t) \end{pmatrix} \tag{1.2}$$

A basic statistical approach consists of considering the $x_i(t)$ as a set of T realizations of m random variables. Thus each set $x_i(t), t = 1, ..., T$ is a sample of one random variable; let us denote the random variable by x_i. In this framework, we could determine the matrix $\mathbf{W}$ by the statistical properties of the transformed components y_i. In the following sections, we discuss some statistical properties that could be used; one of them will lead to independent component analysis.

1.1.2 Dimension reduction methods

One statistical principle for choosing the matrix $\mathbf{W}$ is to limit the number of components y_i to be quite small, maybe only 1 or 2, and to determine $\mathbf{W}$ so that the y_i contain as much information on the data as possible. This leads to a family of techniques called principal component analysis or factor analysis.

In a classic paper, Spearman [409] considered data that consisted of school performance rankings given to schoolchildren in different branches of study, complemented by some laboratory measurements. Spearman then determined $\mathbf{W}$ by finding a single linear combination such that it explained the maximum amount of the variation in the results. He claimed to find a general factor of intelligence, thus founding factor analysis, and at the same time starting a long controversy in psychology.

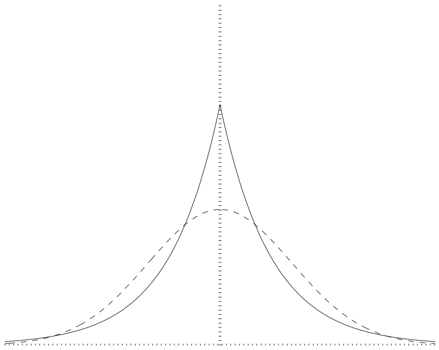

Fig. 1.1 The density function of the Laplacian distribution, which is a typical supergaussian distribution. For comparison, the gaussian density is given by a dashed line. The Laplacian density has a higher peak at zero, and heavier tails. Both densities are normalized to unit variance and have zero mean.

1.1.3 Independence as a guiding principle

Another principle that has been used for determining $\mathbf{W}$ is independence: the components y_i should be statistically independent. This means that the value of any one of the components gives no information on the values of the other components.

In fact, in factor analysis it is often claimed that the factors are independent, but this is only partly true, because factor analysis assumes that the data has a gaussian distribution. If the data is gaussian, it is simple to find components that are independent, because for gaussian data, uncorrelated components are always independent.

In reality, however, the data often does not follow a gaussian distribution, and the situation is not as simple as those methods assume. For example, many real-world data sets have *supergaussian* distributions. This means that the random variables take relatively more often values that are very close to zero or very large. In other words, the probability density of the data is peaked at zero and has heavy tails (large values far from zero), when compared to a gaussian density of the same variance. An example of such a probability density is shown in Fig. 1.1.

This is the starting point of ICA. We want to find *statistically independent* components, in the general case where the data is *nongaussian*.

1.2 BLIND SOURCE SEPARATION

Let us now look at the same problem of finding a good representation, from a different viewpoint. This is a problem in signal processing that also shows the historical background for ICA.

1.2.1 Observing mixtures of unknown signals

Consider a situation where there are a number of signals emitted by some physical objects or sources. These physical sources could be, for example, different brain areas emitting electric signals; people speaking in the same room, thus emitting speech signals; or mobile phones emitting their radio waves. Assume further that there are several sensors or receivers. These sensors are in different positions, so that each records a mixture of the original source signals with slightly different weights.

For the sake of simplicity of exposition, let us say there are three underlying source signals, and also three observed signals. Denote by $x_1(t), x_2(t)$ and $x_3(t)$ the observed signals, which are the amplitudes of the recorded signals at time point t, and by $s_1(t), s_2(t)$ and $s_3(t)$ the original signals. The $x_i(t)$ are then weighted sums of the $s_i(t)$, where the coefficients depend on the distances between the sources and the sensors:

$$\begin{aligned} x_1(t) &= a_{11}s_1(t) + a_{12}s_2(t) + a_{13}s_3(t) \\ x_2(t) &= a_{21}s_1(t) + a_{22}s_2(t) + a_{23}s_3(t) \\ x_3(t) &= a_{31}s_1(t) + a_{32}s_2(t) + a_{33}s_3(t) \end{aligned} \tag{1.3}$$

The a_{ij} are constant coefficients that give the mixing weights. They are assumed *unknown*, since we cannot know the values of a_{ij} without knowing all the properties of the physical mixing system, which can be extremely difficult in general. The source signals s_i are *unknown as well*, since the very problem is that we cannot record them directly.

As an illustration, consider the waveforms in Fig. 1.2. These are three linear mixtures x_i of some original source signals. They look as if they were completely noise, but actually, there are some quite structured underlying source signals hidden in these observed signals.

What we would like to do is to find the original signals from the mixtures $x_1(t), x_2(t)$ and $x_3(t)$. This is the blind source separation (BSS) problem. *Blind* means that we know very little if anything about the original sources.

We can safely assume that the mixing coefficients a_{ij} are different enough to make the matrix that they form invertible. Thus there exists a matrix $\mathbf{W}$ with coefficients w_{ij}, such that we can separate the s_i as

$$\begin{aligned} s_1(t) &= w_{11}x_1(t) + w_{12}x_2(t) + w_{13}x_3(t) \\ s_2(t) &= w_{21}x_1(t) + w_{22}x_2(t) + w_{23}x_3(t) \\ s_3(t) &= w_{31}x_1(t) + w_{32}x_2(t) + w_{33}x_3(t) \end{aligned} \tag{1.4}$$

Such a matrix $\mathbf{W}$ could be found as the inverse of the matrix that consists of the mixing coefficients a_{ij} in Eq. 1.3, if we knew those coefficients a_{ij}.

Now we see that in fact this problem is mathematically similar to the one where we wanted to find a good representation for the random data in $x_i(t)$, as in (1.2). Indeed, we could consider each signal $x_i(t), t = 1, ..., T$ as a sample of a random variable x_i, so that the value of the random variable is given by the amplitudes of that signal at the time points recorded.

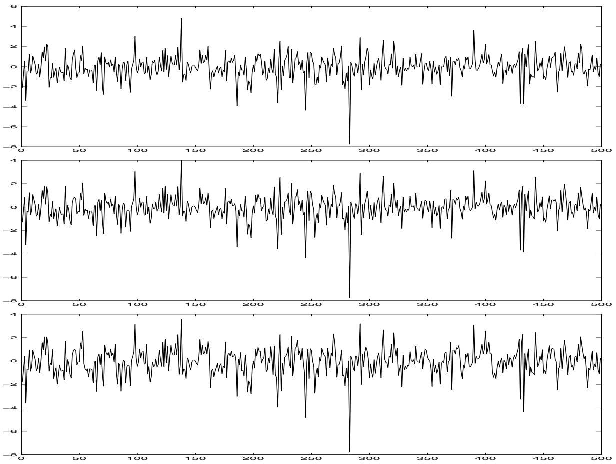

Fig. 1.2 The observed signals that are assumed to be mixtures of some underlying source signals.

1.2.2 Source separation based on independence

The question now is: How can we estimate the coefficients w_{ij} in (1.4)? We want to obtain a general method that works in many different circumstances, and in fact provides one answer to the very general problem that we started with: finding a good representation of multivariate data. Therefore, we use very general statistical properties. All we observe is the signals x_1, x_2 and x_3, and we want to find a matrix $\mathbf{W}$ so that the representation is given by the original source signals s_1, s_2, and s_3.

A surprisingly simple solution to the problem can be found by considering just the statistical independence of the signals. In fact, if the signals are *not gaussian*, it is enough to determine the coefficients w_{ij}, so that the signals

$$\begin{aligned} y_1(t) &= w_{11}x_1(t) + w_{12}x_2(t) + w_{13}x_3(t) \\ y_2(t) &= w_{21}x_1(t) + w_{22}x_2(t) + w_{23}x_3(t) \\ y_3(t) &= w_{31}x_1(t) + w_{32}x_2(t) + w_{33}x_3(t) \end{aligned} \tag{1.5}$$

are statistically independent. If the signals y_1, y_2, and y_3 are independent, then they are equal to the original signals s_1, s_2, and s_3. (They could be multiplied by some scalar constants, though, but this has little significance.)

Using just this information on the statistical independence, we can in fact estimate the coefficient matrix $\mathbf{W}$ for the signals in Fig. 1.2. What we obtain are the source signals in Fig. 1.3. (These signals were estimated by the FastICA algorithm that we shall meet in several chapters of this book.) We see that from a data set that seemed to be just noise, we were able to estimate the original source signals, using an algorithm that used the information on the independence only. These estimated signals are indeed equal to those that were used in creating the mixtures in Fig. 1.2

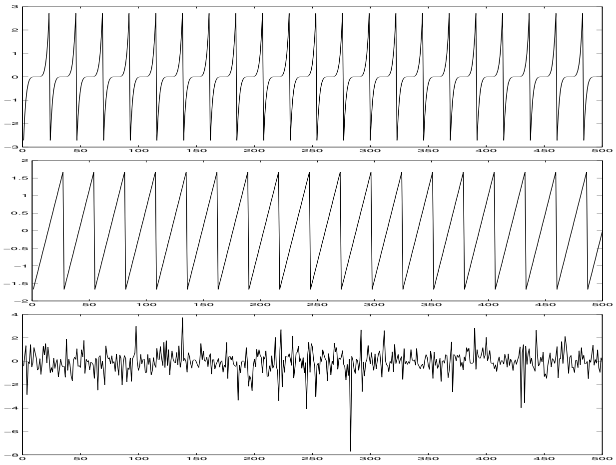

Fig. 1.3 The estimates of the original source signals, estimated using only the observed mixture signals in Fig. 1.2. The original signals were found very accurately.

(the original signals are not shown, but they really are virtually identical to what the algorithm found). Thus, in the source separation problem, the original signals were the "independent components" of the data set.

1.3 INDEPENDENT COMPONENT ANALYSIS

1.3.1 Definition

We have now seen that the problem of blind source separation boils down to finding a linear representation in which the components are statistically independent. In practical situations, we cannot in general find a representation where the components are really independent, but we can at least find components that are as independent as possible.

This leads us to the following definition of ICA, which will be considered in more detail in Chapter 7. Given a set of observations of random variables $(x_1(t), x_2(t), ..., x_n(t))$, where t is the time or sample index, assume that they are generated as a linear mixture of independent components:

$$\begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{pmatrix} = \mathbf{A} \begin{pmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_n(t) \end{pmatrix} \tag{1.6}$$

where $\mathbf{A}$ is some unknown matrix. Independent component analysis now consists of estimating both the matrix $\mathbf{A}$ and the $s_i(t)$, when we only observe the $x_i(t)$. Note that we assumed here that the number of independent components s_i is equal to the number of observed variables; this is a simplifying assumption that is not completely necessary.

Alternatively, we could define ICA as follows: find a linear transformation given by a matrix $\mathbf{W}$ as in (1.2), so that the random variables $y_i, i = 1, ..., n$ are as independent as possible. This formulation is not really very different from the previous one, since after estimating $\mathbf{A}$, its inverse gives $\mathbf{W}$.

It can be shown (see Section 7.5) that the problem is well-defined, that is, the model in (1.6) can be estimated if and only if the components s_i are *nongaussian*. This is a fundamental requirement that also explains the main difference between ICA and factor analysis, in which the nongaussianity of the data is not taken into account. In fact, ICA could be considered as *nongaussian factor analysis*, since in factor analysis, we are also modeling the data as linear mixtures of some underlying factors.

1.3.2 Applications

Due to its generality the ICA model has applications in many different areas, some of which are treated in Part IV. Some examples are:

- In brain imaging, we often have different sources in the brain emit signals that are mixed up in the sensors outside of the head, just like in the basic blind source separation model (Chapter 22).

- In econometrics, we often have parallel time series, and ICA could decompose them into independent components that would give an insight to the structure of the data set (Section 24.1).

- A somewhat different application is in image feature extraction, where we want to find features that are as independent as possible (Chapter 21).

1.3.3 How to find the independent components

It may be very surprising that the independent components can be estimated from linear mixtures with no more assumptions than their independence. Now we will try to explain briefly why and how this is possible; of course, this is the main subject of the book (especially of Part II).

Uncorrelatedness is not enough The first thing to note is that independence is a much stronger property than uncorrelatedness. Considering the blind source separation problem, we could actually find many different uncorrelated representations of the signals that would not be independent and would not separate the sources. Uncorrelatedness in itself is not enough to separate the components. This is also the reason why principal component analysis (PCA) or factor analysis cannot separate the signals: they give components that are uncorrelated, but little more.

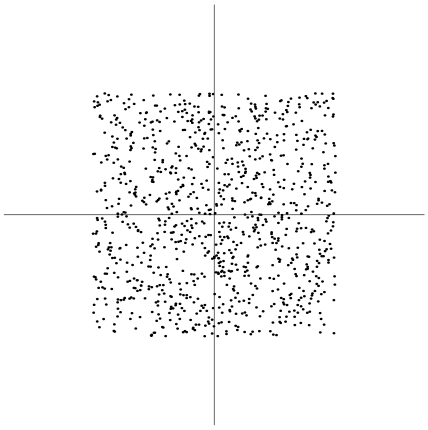

Fig. 1.4 A sample of independent components s_1 and s_2 with uniform distributions. Horizontal axis: s_1; vertical axis: s_2.

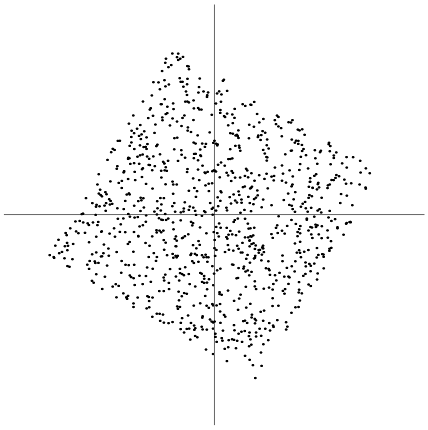

Fig. 1.5 Uncorrelated mixtures x_1 and x_2. Horizontal axis: x_1; vertical axis: x_2.

Let us illustrate this with a simple example using two independent components with uniform distributions, that is, the components can have any values inside a certain interval with equal probability. Data from two such components are plotted in Fig. 1.4. The data is uniformly distributed inside a square due to the independence of the components.

Now, Fig. 1.5 shows two *uncorrelated mixtures* of those independent components. Although the mixtures are uncorrelated, one sees clearly that the distributions are not the same. The independent components are still mixed, using an orthogonal mixing matrix, which corresponds to a rotation of the plane. One can also see that in Fig. 1.5 the components are not independent: if the component on the horizontal axis has a value that is near the corner of the square that is in the extreme right, this clearly restricts the possible values that the components on the vertical axis can have.

In fact, by using the well-known decorrelation methods, we can transform any linear mixture of the independent components into uncorrelated components, in which case the mixing is orthogonal (this will be proven in Section 7.4.2). Thus, the trick in ICA is to estimate the orthogonal transformation that is left after decorrelation. This is something that classic methods cannot estimate because they are based on essentially the same covariance information as decorrelation.

Figure 1.5 also gives a hint as to why ICA is possible. By locating the edges of the square, we could compute the rotation that gives the original components. In the following, we consider a couple more sophisticated methods for estimating ICA.

Nonlinear decorrelation is the basic ICA method One way of stating how independence is stronger than uncorrelatedness is to say that independence implies *nonlinear uncorrelatedness*: If s_1 and s_2 are independent, then any nonlinear transformations $g(s_1)$ and $h(s_2)$ are uncorrelated (in the sense that their covariance is zero). In contrast, for two random variables that are merely uncorrelated, such nonlinear transformations do not have zero covariance in general.

Thus, we could attempt to perform ICA by a stronger form of decorrelation, by finding a representation where the y_i are uncorrelated even after some nonlinear transformations. This gives a simple principle of estimating the matrix $\mathbf{W}$:

> **ICA estimation principle 1**: Nonlinear decorrelation. Find the matrix $\mathbf{W}$ so that for any $i \neq j$, the components y_i and y_j are uncorrelated, *and* the transformed components $g(y_i)$ and $h(y_j)$ are uncorrelated, where g and h are some suitable nonlinear functions.

This is a valid approach to estimating ICA: If the nonlinearities are properly chosen, the method does find the independent components. In fact, computing nonlinear correlations between the two mixtures in Fig. 1.5, one would immediately see that the mixtures are not independent.

Although this principle is very intuitive, it leaves open an important question: How should the nonlinearities g and h be chosen? Answers to this question can be found be using principles from estimation theory and information theory. Estimation theory provides the most classic method of estimating any statistical model: the *maximum likelihood* method (Chapter 9). Information theory provides exact measures of independence, such as *mutual information* (Chapter 10). Using either one of these theories, we can determine the nonlinear functions g and h in a satisfactory way.

Independent components are the maximally nongaussian components

Another very intuitive and important principle of ICA estimation is maximum nongaussianity (Chapter 8). The idea is that according to the central limit theorem, sums of nongaussian random variables are closer to gaussian that the original ones. Therefore, if we take a linear combination $y = \sum_i b_i x_i$ of the observed mixture variables (which, because of the linear mixing model, is a linear combination of the independent components as well), this will be maximally nongaussian if it equals one of the independent components. This is because if it were a real mixture of two or more components, it would be closer to a gaussian distribution, due to the central limit theorem.

Thus, the principle can be stated as follows

> **ICA estimation principle 2**: Maximum nongaussianity. Find the local maxima of nongaussianity of a linear combination $y = \sum_i b_i x_i$ under the constraint that the variance of y is constant. Each local maximum gives one independent component.

To measure nongaussianity in practice, we could use, for example, the *kurtosis*. Kurtosis is a higher-order cumulant, which are some kind of generalizations of variance using higher-order polynomials. Cumulants have interesting algebraic and statistical properties which is why they have an important part in the theory of ICA.

For example, comparing the nongaussianities of the components given by the axes in Figs. 1.4 and 1.5, we see that in Fig. 1.5 they are smaller, and thus Fig. 1.5 cannot give the independent components (see Chapter 8).

An interesting point is that this principle of maximum nongaussianity shows the very close connection between ICA and an independently developed technique called *projection pursuit*. In projection pursuit, we are actually looking for maximally nongaussian linear combinations, which are used for visualization and other purposes. Thus, the independent components can be interpreted as projection pursuit directions.

When ICA is used to extract features, this principle of maximum nongaussianity also shows an important connection to *sparse coding* that has been used in neuroscientific theories of feature extraction (Chapter 21). The idea in sparse coding is to represent data with components so that only a small number of them are "active" at the same time. It turns out that this is equivalent, in some situations, to finding components that are maximally nongaussian.

The projection pursuit and sparse coding connections are related to a deep result that says that ICA gives a linear representation that is *as structured as possible*. This statement can be given a rigorous meaning by information-theoretic concepts (Chapter 10), and shows that the independent components are in many ways easier to process than the original random variables. In particular, independent components are easier to code (compress) than the original variables.

ICA estimation needs more than covariances There are many other methods for estimating the ICA model as well. Many of them will be treated in this book. What they all have in common is that they consider some statistics that are not contained in the covariance matrix (the matrix that contains the covariances between all pairs of the x_i).

Using the covariance matrix, we can decorrelate the components in the ordinary linear sense, but not any stronger. Thus, all the ICA methods use some form of *higher-order statistics*, which specifically means information not contained in the covariance matrix. Earlier, we encountered two kinds of higher-order information: the nonlinear correlations and kurtosis. Many other kinds can be used as well.

Numerical methods are important In addition to the estimation principle, one has to find an algorithm for implementing the computations needed. Because the estimation principles use nonquadratic functions, the computations needed usually cannot be expressed using simple linear algebra, and therefore they can be quite demanding. Numerical algorithms are thus an integral part of ICA estimation methods.

The numerical methods are typically based on optimization of some objective functions. The basic optimization method is the gradient method. Of particular interest is a fixed-point algorithm called FastICA that has been tailored to exploit the particular structure of the ICA problem. For example, we could use both of these methods to find the maxima of the nongaussianity as measured by the absolute value of kurtosis.

1.4 HISTORY OF ICA

The technique of ICA, although not yet the name, was introduced in the early 1980s by J. Hérault, C. Jutten, and B. Ans [178, 179, 16]. As recently reviewed by Jutten [227], the problem first came up in 1982 in a neurophysiological setting. In a simplified model of motion coding in muscle contraction, the outputs $x_1(t)$and $x_2(t)$ were two types of sensory signals measuring muscle contraction, and $s_1(t)$ and $s_2(t)$ were the angular position and velocity of a moving joint. Then it is not unreasonable to assume that the ICA model holds between these signals. The nervous system must be somehow able to infer the position and velocity signals $s_1(t), s_2(t)$ from the measured responses $x_1(t), x_2(t)$. One possibility for this is to learn the inverse model using the nonlinear decorrelation principle in a simple neural network. Hérault and Jutten proposed a specific feedback circuit to solve the problem. This approach is covered in Chapter 12.

All through the 1980s, ICA was mostly known among French researchers, with limited influence internationally. The few ICA presentations in international neural network conferences in the mid-1980s were largely buried under the deluge of interest in back-propagation, Hopfield networks, and Kohonen's Self-Organizing Map (SOM), which were actively propagated in those times. Another related field was higher-order spectral analysis, on which the first international workshop was organized in 1989. In this workshop, early papers on ICA by J.-F. Cardoso [60] and P. Comon [88] were given. Cardoso used algebraic methods, especially higher-order cumulant tensors, which eventually led to the JADE algorithm [72]. The use of fourth-order cumulants has been earlier proposed by J.-L. Lacoume [254]. In signal processing literature, classic early papers by the French group are [228, 93, 408, 89]. A good source with historical accounts and a more complete list of references is [227].

In signal processing, there had been earlier approaches in the related problem of blind signal deconvolution [114, 398]. In particular, the results used in multichannel blind deconvolution are very similar to ICA techniques.

The work of the scientists in the 1980's was extended by, among others, A. Cichocki and R. Unbehauen, who were the first to propose one of the presently most popular ICA algorithms [82, 85, 84]. Some other papers on ICA and signal separation from early 1990s are [57, 314]. The "nonlinear PCA" approach was introduced by the present authors in [332, 232]. However, until the mid-1990s, ICA remained a rather small and narrow research effort. Several algorithms were proposed that worked, usually in somewhat restricted problems, but it was not until later that the rigorous connections of these to statistical optimization criteria were exposed.

ICA attained wider attention and growing interest after A.J. Bell and T.J. Sejnowski published their approach based on the infomax principle [35, 36] in the mid-90's. This algorithm was further refined by S.-I. Amari and his co-workers using the natural gradient [12], and its fundamental connections to maximum likelihood estimation, as well as to the Cichocki-Unbehauen algorithm, were established. A couple of years later, the present authors presented the fixed-point or FastICA algorithm, [210, 192, 197], which has contributed to the application of ICA to large-scale problems due to its computational efficiency.

Since the mid-1990s, there has been a growing wave of papers, workshops, and special sessions devoted to ICA. The first international workshop on ICA was held in Aussois, France, in January 1999, and the second workshop followed in June 2000 in Helsinki, Finland. Both gathered more than 100 researchers working on ICA and blind signal separation, and contributed to the transformation of ICA to an established and mature field of research.

