Independent Component Analysis

James V. Stone

November 14, 2014

Sheffield University, Sheffield, UK
[1]

Keywords: ***independent component analysis, independence, blind source separation, projection pursuit, complexity pursuit***

Abstract

Abstract: Given a set of M signal mixtures $(x^1, x^2, \ldots, x^M)$ (e.g. microphone outputs), each of which is a different mixture of a set of M statistically independent source signals $(s^1, s^2, \ldots, s^M)$ (e.g. voices), independent component analysis (ICA) recovers the source signals (voices) from the signal mixtures. ICA is based on the assumptions that source signals are statistically independent and that they have non-Gaussian distributions. Different physical processes usually generate statistically independent and non-Gaussian signals, so that, in the process of extracting such signals from a set of signal mixtures, ICA effectively recovers the underlying physical causes for a given set of measured signal mixtures.

Introduction

Most measured quantities are actually mixtures of other quantities. Typical examples are (a) the sound in a room in which several people are talking simultaneously, (b) an electroencephalogram (EEG) signal, which contains contributions from many different brain regions, and (c) a person's height, which is determined by contributions from many different genetic and environmental factors. Science is, to a large extent, concerned with establishing the precise nature of the component processes responsible for a given set of measured quantities, whether these involve EEG signals, human height, or even IQ. Under certain conditions, the signals underlying measured quantities can be recovered by making use of *independent component analysis* (ICA), which is a member of a class of *blind source separation* (BSS) methods.

The success of ICA depends on a single highly plausible assumption regarding the nature of the physical world: independent variables or signals[1] are generated by different underlying

[1]This article was originally published online in 2005 in Encyclopedia of Statistics in Behavioral Science,

physical processes. If two signals are independent, then the value of one signal cannot be used to predict anything about the other signal. In practice, most measured signals are derived from many independent physical processes, and are therefore mixtures of independent signals. Given such a set of measured signals (i.e., mixtures), ICA works by finding a transformation of those mixtures, which produces independent signal components, on the assumption that each of these independent component signals is associated with a different physical process. In the language of ICA, the measured signals are known as *signal mixtures*, and the required independent signals are known as *source signals*.

ICA has been applied to separation of different speech signals [1][2], analysis of EEG data [7], functional magnetic resonance imaging (fMRI) data [8], **image processing** [2], and as a model of biological image processing [11]. A review of recent advances in ICA can be found in [6].

Before embarking on an account of the mathematical details of ICA, a simple, intuitive example of how ICA could separate two speech signals is given. However, it should be noted that this example could equally well apply to *any physically measured set of signals, and to any number of signals* (e.g., images, biomedical data, or stock prices).

Applying ICA to Speech Data

Consider two people speaking at the same time in a room containing two microphones, as depicted in Figure 1. If each voice signal is examined at a fine time scale, then it is apparent that the amplitude of one voice at any given point in time is unrelated to the amplitude of the other voice at that time. The reason that the amplitudes of two voices are unrelated is that they are generated by two unrelated physical processes (i.e., by two different people). If we know that the voices are unrelated, then one key strategy for separating voice mixtures (e.g., microphone outputs) into their constituent voice components is to extract unrelated time-varying signals from these mixtures. The property of being unrelated is of fundamental importance.

While it is true that two voice signals are unrelated, this informal observation can be defined formally in terms of *statistical independence*, which is often truncated to *independence*. If two or more signals are statistically independent of each other, then the value of one signal provides no information regarding the value of the other signals.

The Number of Sources and Mixtures

One important fact about ICA is often not appreciated. Basically, there must usually be at least as many different mixtures of a set of source signals as there are source signals (but see [10]). For the example of speech signals, this implies that there must be at least as many microphones (different voice mixtures) as there are voices (source signals).

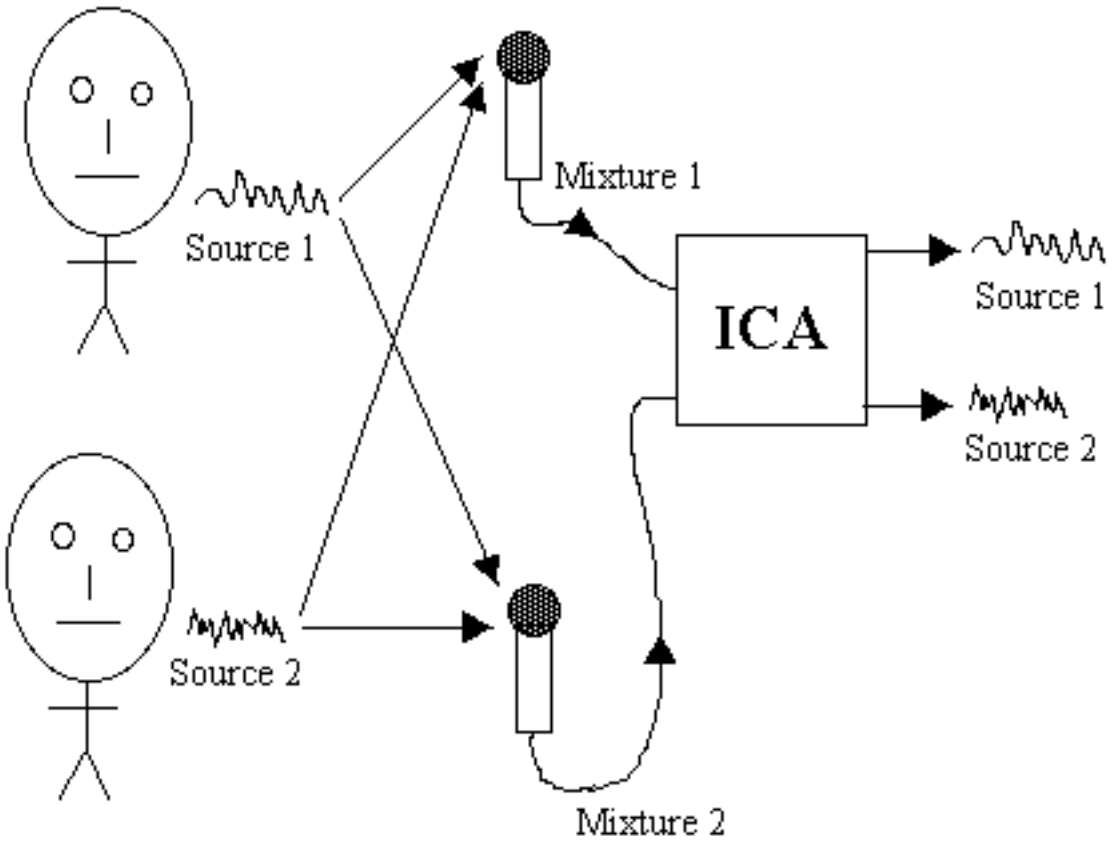

Figure 1: ICA in a nutshell: If two people speak at the same time in a room containing two microphones, then the output of each microphone is a *mixture* of two voice signals. Given these two *signal mixtures*, ICA can recover the two original voices or *source signals*. This example uses speech, but ICA can extract source signals from any set of two or more measured signal mixtures, where each signal mixture is assumed to consist of a linear mixture of source signals (see section 'Mixing Signals')

Effects of Mixing Signals

When a set of two or more independent source signals are mixed to make a corresponding set of signal mixtures, as shown in Figure 1, three effects follow.

- *Independence.* Whereas source signals are independent, their signal mixtures are not. This is because each source signal contributes to every mixture, and the mixtures cannot, therefore, be independent.
- *Normality.* The central limit theorem ensures that a signal mixture that is the sum of almost any signals yields a bell-shaped, **normal** or *Gaussian* histogram. In contrast, the histogram of a typical source signal has a non-Gaussian structure (see Figure 2).
- *Complexity.* The complexity of any mixture is greater than (or equal to) that of its simplest (i.e., least complex) constituent source signal. This ensures that extracting the least complex signal from a set of signal mixtures yields a source signal [10].

These three effects can be used either on their own or in combination to extract source signals from signal mixtures. The effects labeled *normality* and *complexity* are used in **projection pursuit** [5] and *complexity pursuit* [4, 9], respectively, and the effects labeled *independence* and *normality* are used together in ICA (also see [10]).

Representing Multiple Signals

A speech source signal s_1 is represented as $s_1 = (s_1^1, s_1^2, \ldots, s_1^N)$, where s_1 adopts amplitudess_1^1, then s_1^2, and so on; superscripts specify time and subscripts specify signal identity (e.g., speaker identity). We will be considering how to mix and unmix a *set* of two or more signals, and we define a specific set of two time-varying speech signals s_1 and s_2 in order to provide a concrete example. Now, the amplitudes of both signals can be written as a *vector variable* $\mathbf{s}$, which can be rewritten in one of several mathematically equivalent forms:

$$\mathbf{s} = \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} \tag{1}$$

$$= \begin{pmatrix} (s_1^1, & s_1^2, & \ldots, & s_1^N) \\ (s_2^1, & s_2^2, & \ldots, & s_2^N) \end{pmatrix}. \tag{2}$$

We introduce the transpose operator, which simply transforms rows into columns (or vice versa), and is defined by $\mathbf{s} = (s_1, s_2)^T$.

Mixing Signals

The different distance of each source (i.e., person) from each microphone ensures that each source contributes a different amount to the microphone's output. The microphone's output

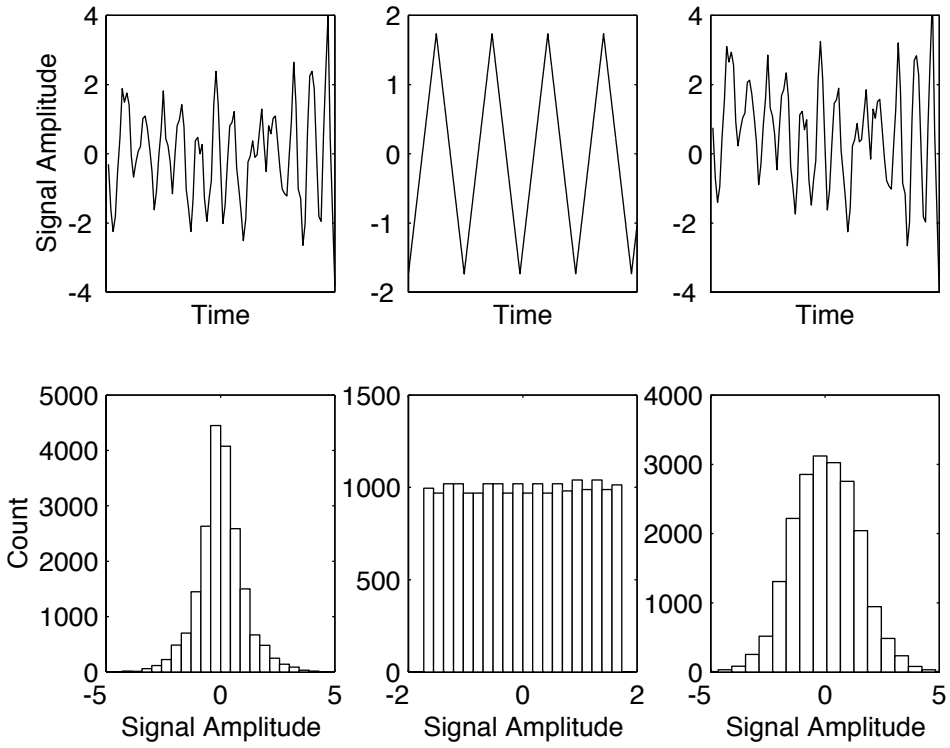

Figure 2: Signal mixtures have Gaussian or normal histograms. Signals (top row) and corresponding histograms of signal values (bottom row), where each histogram approximates the *probability density function* (pdf) of one signal. The top panels display only a small segment of the signals used to construct displayed histograms. A speech source signal (a), and a histogram of amplitude values in that signal (d). A sawtooth source signal (b), and its histogram (e). A signal mixture (c), which is the sum of the source signals on the left and middle, and its bell-shaped histogram (f) [fg002.eps]

is, therefore, a *linear* mixture x_1 that consists of a weighted sum of the two source signals $x_1 = as_1 + bs_2$, where the *mixing coefficients* a and b are determined by the distance of each source from each microphone. As we are concerned here with unmixing a set of two signal mixtures (see Figure 1), we need another microphone in a different location from the first. In this case, the microphone's output x_2 is $x_2 = cs_1 + ds_2$, where the mixing coefficients are c and d.

Unmixing Signals

Generating mixtures from source signals in this linear manner ensures that each source signal can be recovered by linearly recombining signal mixtures. The precise nature of this recombination is determined by a set of *unmixing coefficients* (α, β, γ, δ), such that $s_1 = \alpha x_1 + \beta x_2$ and $s_2 = \gamma x_1 + \delta x_2$. Thus, the problem solved by ICA, and by all other *BSS* methods, consists of finding values for these unmixing coefficients.

The Mixing and Unmixing Matrices

The set of mixtures defines a vector variable $\mathbf{x} = (x_1, x_2)^T$, and the transformation from $\mathbf{s}$ to $\mathbf{x}$ defines a *mixing matrix* $\mathbf{A}$:

$$\begin{aligned}\mathbf{x} &= \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} s_1^1, & s_1^2, & \ldots, & s_1^N \\ s_2^1, & s_2^2, & \ldots, & s_2^N \end{pmatrix} \\ &= \mathbf{A}\mathbf{s}.\end{aligned} \tag{3}$$

The mapping from $\mathbf{x}$ to $\mathbf{s} = (s_1, s_2)^T$ defines an optimal *unmixing matrix* $\mathbf{W}^* = (\mathbf{w}_1, \mathbf{w}_2)^T$ with (row) weight vectors $\mathbf{w}_1^T = (\alpha, \beta)$ and $\mathbf{w}_2^T = (\gamma, \delta)$

$$\begin{aligned}\mathbf{s} &= \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \begin{pmatrix} x_1^1, & x_1^2, & \ldots, & x_1^N \\ x_2^1, & x_2^2, & \ldots, & x_2^N \end{pmatrix} \\ &= (\mathbf{w}_1, \mathbf{w}_2)^T (x_1, x_2)\end{aligned} \tag{4}$$

$$= \mathbf{W}^*\mathbf{x}. \tag{5}$$

It can be seen that $\mathbf{W}^*$ reverses, or inverts, the effects of $\mathbf{A}$, and indeed, $\mathbf{W}^*$could be estimated from the *matrix inverse* $\mathbf{W}^* = \mathbf{A}^{-1}$, if $\mathbf{A}$ were known[3]. However, as we are ultimately concerned with finding $\mathbf{W}^*$ when $\mathbf{A}$ is not known, we cannot, therefore, use $\mathbf{A}^{-1}$ to estimate $\mathbf{W}^*$. For arbitrary values of the unmixing coefficients, the unmixing matrix is suboptimal and is denoted $\mathbf{W}$. In this case, the signals extracted by $\mathbf{W}$ are not necessarily source signals, and are denoted $\mathbf{y} = \mathbf{W}\mathbf{x}$.

Maximum Likelihood ICA

In practice, it is extremely difficult to measure the independence of a set of extracted signals unless we have some general knowledge about those signals. In fact, the observations above

suggest that we do often have some knowledge of the source signals. Specifically, we know that they are non-Gaussian, and that they are independent. This knowledge can be specified in terms of a formal model, and we can then extract signals that conform to this model. More specifically, we can search for an unmixing matrix that maximizes the agreement between the model and the signals extracted by that unmixing matrix.

One common interpretation of ICA is as a **maximum likelihood** (ML) method for estimating the optimal unmixing matrix $\mathbf{W}^*$. Maximum likelihood estimation (MLE) is a standard statistical tool for finding model parameter values (e.g., the unmixing matrix $\mathbf{W}$) which provide the best fit of a model to a given data set (e.g., the mixtures $\mathbf{x}$). The ICA ML model includes the adjustable parameters in $\mathbf{W}$, and a (usually fixed) model of the source signals. However, this source signal model is quite vague because it is specified only in terms of the general shape of the histogram of source signals. The fact that the source signal model is vague is desirable, because it means that we do not have to know very much about the source signals in order to extract them from the signal mixtures. In essence, a vague source signal model corresponds to a principle of least commitment with respect to the assumed form of the source signals.

As noted above, mixtures of source signals are almost always Gaussian (see Figure 2), and it is fairly safe to assume that non-Gaussian signals must, therefore, be source signals. The amount of 'Gaussian-ness' of a signal can be specified in terms of a histogram of its values, which is an approximation to the **probability density function** (pdf) of that signal (see Figure 2), which is represented as $p_s(s)$. The value of the function $p_s(s^t)$ is the *probability density* of the signal s at $s = s^t$, where t represents a particular point in time. Because a pure speech signal contains a high proportion of silence, its pdf is highly 'peaky' or **leptokurtotic**, with a peak around zero (see Figure 3). It, therefore, makes sense to specify a leptokurtotic function as our model pdf for speech source signals.

We know the source signals are independent, and we need to incorporate this knowledge into our model. The degree of mutual independence between signals can be specified in terms of their *joint pdf* (see Figure 3). By analogy, with the pdf of a scalar signal, a joint pdf defines the probability that the values of a set of signals $\mathbf{s} = (s_1, s_2)^T$ fall within a small range around a specific set of values $\mathbf{s}^t = (s_1^t, s_2^t)^T$. Crucially, if these signals are mutually independent, then the joint pdf $p_\mathbf{s}$ of $\mathbf{s}$ can be expressed as the product of the pdfs (p_{s_1}, p_{s_2}) of its constituent signals s_1 and s_2. That is, $p_\mathbf{s}(\mathbf{s}) = p_{s_1}(s_1) \times p_{s_2}(s_2)$, where the pdfs $p_{s_1}(s_1)$ and $p_{s_2}(s_2)$ of the signals s_1 and s_2 (respectively) are the *marginal* pdfs of the joint pdf $p_\mathbf{s}$.

Using ML ICA, *the objective is to find an unmixing matrix* $\mathbf{W}$ *that yields extracted signals* $\mathbf{y} = \mathbf{W}\mathbf{x}$, *which have a joint pdf as similar as possible to the model joint pdf* $p_\mathbf{s}(\mathbf{s})$ *of the unknown source signals* $\mathbf{s}$. This model incorporates the assumptions that source signals are non-Gaussian (leptokurtotic, in the case of speech) and independent. Fortunately, ICA seems to be very robust with respect to differences between model pdfs and the pdfs of source signals [3]. In other words, ICA is tolerant with respect to violations of the assumptions on which it is based. Note that, because $\mathbf{A}$ and $\mathbf{W}$ are inverses of each other[5], it does not matter whether the model parameters are expressed in terms of $\mathbf{A}$ or$\mathbf{W}$.

When using ML ICA, we consider the probability of obtaining the observed mixtures $\mathbf{x}$

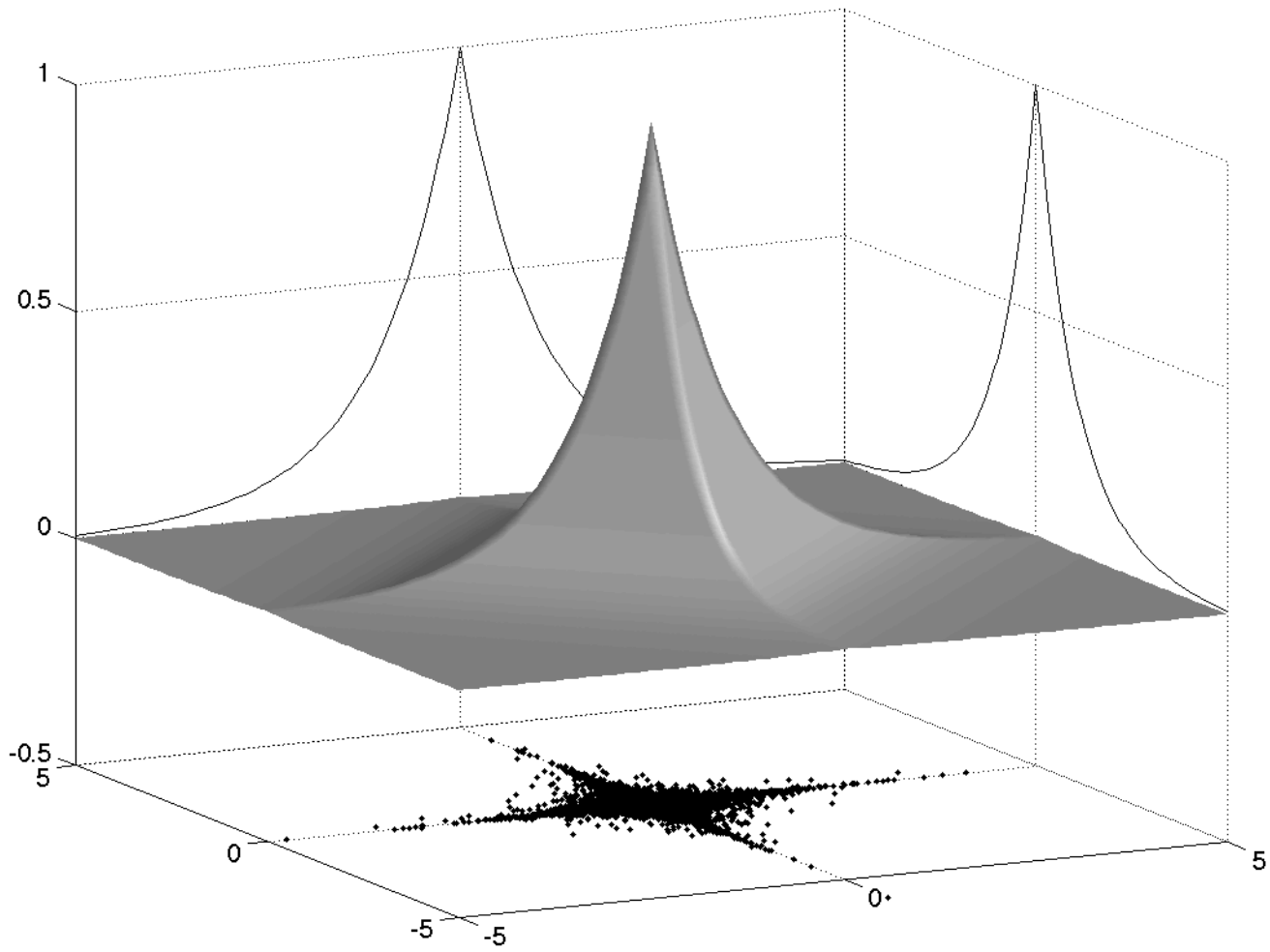

Figure 3: Marginal and joint probability density function (pdfs) of two high-kurtosis independent variables (e.g., speech signals). Given a set of signals $\mathbf{s} = (s_1, s_2)^T$, the pdf of each signal is essentially a histogram of values in that signal, as indicated by the two curves plotted along the horizontal axes. Similarly, the joint pdf $p_{\mathbf{s}}(\mathbf{s})$ of two signals is essentially a two-dimensional histogram of pairs of signal values $\mathbf{s}^t = (s_1^t, s_2^t)$ at time t. Accordingly, the joint probability of observing values $\mathbf{s}^t = (s_1^t, s_2^t)$ is indicated by the local density of plotted points on the horizontal plane. This local density is an approximation to the joint pdf $p_{\mathbf{s}}(\mathbf{s})$, which is indicated by the height of the solid surface. The pdfs $p_{s_1}(s_1)$ and $p_{s_2}(s_2)$ of the signals $s_{\mathbf{1}}$ and $s_{\mathbf{2}}$ are the *marginal* pdfs of the joint pdf $p_{\mathbf{s}}(\mathbf{s})$.

given putative model parameter values $\mathbf{A}$, where this probability is known as the *likelihood* of the model parameter values $\mathbf{A}$. We can then pose the question: given that the source signals have a joint pdf$p_{\mathbf{s}}(\mathbf{s})$, which particular mixing matrix $\mathbf{A}$ (and, therefore, which unmixing matrix $\mathbf{W} = \mathbf{A}^{-1}$) is most likely to have generated the observed signal mixtures $\mathbf{x}$? In other words, if the probability of obtaining the observed mixtures (from some unknown source signals with joint pdf $p_{\mathbf{s}}(\mathbf{s})$) were to vary with $\mathbf{A}$, then which particular $\mathbf{A}$ would maximize this probability?

MLE is based on the assumption that if the model joint pdf and the model parameters $\mathbf{A}$ are correct, then a high probability should be obtained for the mixtures $\mathbf{x}$ that were actually observed. Conversely, if $\mathbf{A}$ is far from the correct parameter values, then a low probability of the observed mixtures would be expected. We will assume that all source signals have the same (leptokurtotic) pdf $p_s(\mathbf{s})$. This may not seem much to go on, but it turns out to be perfectly adequate for extracting source signals from signal mixtures.

The Nuts and Bolts of ML ICA

Consider a (mixture) vector variable $\mathbf{x}$ with joint pdf $p_{\mathbf{x}}(\mathbf{x})$, and a (source) vector variable $\mathbf{s}$ with joint pdf$p_{\mathbf{s}}(\mathbf{s})$, such that $\mathbf{s} = \mathbf{W}^*\mathbf{x}$, where $\mathbf{W}^*$ is the optimal unmixing matrix. As noted above, the number of source signals and mixtures must be equal, which ensures that the matrix $\mathbf{W}^*$ is square. In general, the relation between the joint pdfs of $\mathbf{x}$ and $\mathbf{s}$ is

$$p_{\mathbf{x}}(\mathbf{x}) = p_{\mathbf{s}}(\mathbf{s})|\mathbf{W}^*|, \tag{6}$$

where $|\mathbf{W}^*| = |\partial\mathbf{s}/\partial\mathbf{x}|$ is the *Jacobian* of $\mathbf{s}$ with respect to $\mathbf{x}$. Equation (6) defines the **likelihood** of the unmixing matrix $\mathbf{W}^*$, which is the probability of $\mathbf{x}$ given$\mathbf{W}^*$.

Given an unmixing matrix $\mathbf{W}$, the extracted signals are $\mathbf{y} = \mathbf{W}\mathbf{x}$. Making the dependence on $\mathbf{W}$ explicit, the probability $p_{\mathbf{x}}(\mathbf{x}|\mathbf{W})$ of the signal mixtures $\mathbf{x}$ given $\mathbf{W}$ is

$$p_{\mathbf{x}}(\mathbf{x}|\mathbf{W}) = p_{\mathbf{s}}(\mathbf{W}\mathbf{x})|\mathbf{W}|. \tag{7}$$

We would naturally expect $p_{\mathbf{x}}(\mathbf{x}|\mathbf{W})$ to be maximal if $\mathbf{W} = \mathbf{W}^*$. Thus, (7) can be used to evaluate the quality of any putative unmixing matrix $\mathbf{W}$ in order to find that particular $\mathbf{W}$ which maximizes $p_{\mathbf{x}}(\mathbf{x}|\mathbf{W})$. By convention, (7) defines a *likelihood function* $L(\mathbf{W})$ of $\mathbf{W}$, and its logarithm defines the *log likelihood function* $\ln L(\mathbf{W})$. If the M source signals are mutually independent, so that the joint pdf $p_{\mathbf{s}}(\mathbf{s})$ is the product of its M marginal pdfs, then (7) can be written

$$\ln L(\mathbf{W}) = \sum_{i}^{M} \sum_{t}^{N} \ln p_s(\mathbf{w}_i^T \mathbf{x}_t) + N \ln |\mathbf{W}|. \tag{8}$$

Note that the likelihood $L(\mathbf{W})$ is the probability $p_{\mathbf{x}}(\mathbf{x}|\mathbf{W})$ of $\mathbf{x}$, but using MLE, it is treated as if it were a function of the parameters $\mathbf{W}$. If we substitute a commonly used leptokurtotic model joint pdf for the source signals $p_{\mathbf{s}}(\mathbf{y}) = (1 - \tanh(\mathbf{y})^2)$, then

$$\ln L(\mathbf{W}) = \sum_{i}^{M} \sum_{t}^{N} \ln(1 - \tanh(\mathbf{w}_i^T \mathbf{x}_t)^2) + N \ln |\mathbf{W}|. \tag{9}$$

The matrix $\mathbf{W}$ which maximizes this function is the *maximum likelihood estimate* (MLE) of the optimal unmixing matrix $\mathbf{W}^*$. Equation (9) provides a measure of similarity between the joint pdf of the extracted signals $\mathbf{y} = \mathbf{Wx}$ and the joint model pdf of the source signals $\mathbf{s}$. Having such a measure permits us to use standard optimization methods to iteratively update the unmixing matrix in order to maximize this measure of similarity.

ICA, Principal Component Analysis and Factor Analysis

ICA is related to conventional methods for analyzing large data sets such as principal component analysis (PCA) and factor analysis (FA). Whereas ICA finds a set of source signals which are mutually independent, PCA and FA find a set of signals that are mutually decorrelated (consequently, neither PCA nor FA could extract speech signals, for example). The 'forward' assumption that signals from different physical processes are uncorrelated still holds, but the 'reverse' assumption that uncorrelated signals are from different physical processes does not. This is because lack of correlation is a weaker property than independence. In summary, independence implies a lack of correlation, but a lack of correlation does not imply independence.

Notes

[1] We use the term signal and variable interchangeably here.

[2] This is a seminal paper, which initiated the recent interest in ICA.

[3] The matrix inverse is analogous to the more familiar inverse for scalar variables, such as $x^{-1} = 1/x$.

[4] Up to an irrelevant permutation of rows.

References

[1] Bell, A. J. & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution, *Neural Computation* **7**, 1129 – 1159.

[2] Bell, A. J. & Sejnowski, T. J. (1997). The independent components of natural scenes are edge filters, *Vision Research* **37**(23), 3327 – 3338.

[3] Cardoso, J. (2000). On the stability of source separation algorithms, *Journal of VLSI Signal Processing Systems* **26**(1/2), 7 – 14.

[4] Hyvärinen, A. (2001). Complexity pursuit: separating interesting components from time series, *Neural Computation* **13**, 883 – 898.

[5] Hyvärinen, A., Karhunen, J. & Oja, E. (2001). *Independent Component Analysis*, John Wiley & Sons, New York.

[6] Hyvärinen, A., (2013). *Independent component analysis: recent advances*, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **371**(1984), 1471 – 2962.

[7] Makeig, S., Jung, T., Bell, A. J., Ghahremani, D. & Sejnowski, T. J. (1997). Blind separation of auditory event-related brain responses into independent components, *Proceedings National Academy of Sciences of the United States of America* **94**, 10979 – 10984.

[8] McKeown, M. J., Makeig, S., Brown, G. G., Jung, T. P., Kindermann, S. S. & Sejnowski, T. J. (1998). Spatially independent activity patterns in functional magnetic resonance imaging data during the stroop color-naming task, *Proceedings National Academy of Sciences of the United States of America* **95**, 803 – 810.

[9] Stone, J. V. (2001). Blind source separation using temporal predictability, *Neural Computation* **13**(7), 1559 – 1574.

[10] Stone, J. V. (2004). *Independent Component Analysis: A Tutorial Introduction*, MIT Press, Boston.

[11] van Hateren, J. H. & van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex, *Proceedings of the Royal Society of London. Series B. Biological Sciences* **265**(7), 359 – 366.