1

Introduction

S.J. Roberts

R.M. Everson

1.1 Introduction

Independent Component Analysis (ICA) has recently become an important tool for modelling and understanding empirical datasets as it offers an elegant and practical methodology for *blind* source separation and deconvolution. It is seldom possible to observe a pure unadulterated signal. Instead most observations consist of a mixture of signals usually corrupted by noise, and frequently filtered. The signal processing community has devoted much attention to the problem of recovering the constituent sources from the convolutive mixture; ICA may be applied to this Blind Source Separation (BSS) problem to recover the sources. As the appellation *independent* suggests, recovery relies on the assumption that the constituent sources are mutually independent.

Finding a natural coordinate system is an essential first step in the analysis of empirical data. Principal component analysis (PCA) has, for many years, been used to find a set of basis vectors which are determined by the dataset itself. The principal components are orthogonal and projections of the data onto them are linearly decorrelated, properties which can be ensured by considering only the second order statistical characteristics of the data. ICA aims at a loftier goal: it seeks a transformation to coordinates in which the data are maximally statistically independent, not merely decorrelated.

Perhaps the most famous illustration of ICA is the 'cocktail party problem', in which a listener is faced with the problem of separating the independent voices chattering at a cocktail party. Humans employ many different strategies, often concentrating on just one voice, more or less successfully [Bregman, 1990]. The computational problem of separating

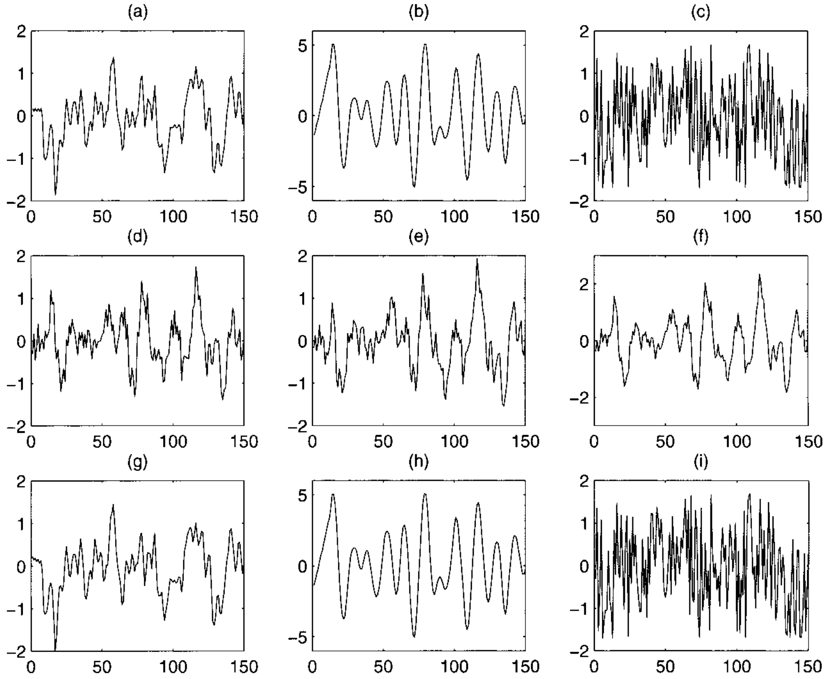

Figure 1.1. ***Mixing and separation of music and noise.*** *Top row: 150 samples of the original sources, $s_m(t)$ ($f_{samp} = 11.3\,kHz$). Middle row: mixtures of the sources,* $\mathbf{x}(t)$. *Bottom row: the estimated sources, $a_m(t)$. To facilitate comparison both the sources and the recovered sources have been normalised to unit variance.*

the speakers from audio mixtures recorded by microphones is challenging, especially when echoes and time delays are taken into account.

To make the ideas and notation concrete we consider a simple example with three sources. The sources were two fragments of music (a Beethoven string quartet and an old recording of a Bessie Smith blues ballad) and uniform noise. Writing the source signals at the instant t in vector form, $\mathbf{s}(t) = [s_1(t), s_2(t), s_3(t)]^{\mathsf{T}}$, observations $\mathbf{x}(t) \in \mathbb{R}^3$ were generated by mixing the sources by a *mixing matrix*, A, whose elements were chosen at random:†

$$\mathbf{x}(t) = A\mathbf{s}(t). \tag{1.1}$$

The top and middle rows of figure 1.1 show 150 samples from original sources, $\mathbf{s}(t)$, and the mixture $\mathbf{x}(t)$. The aim of BSS is to recover the

† $A = \begin{bmatrix} 0.2519 & 0.0513 & 0.0771 \\ 0.5174 & 0.6309 & 0.4572 \\ 0.1225 & 0.6074 & 0.4971 \end{bmatrix}$

original sources from the observations alone, without any additional knowledge of the sources or their characteristics. Independent component analysis accomplishes the separation relying on the assumption that the sources are independent. It seeks a *separating matrix* (or *filter matrix*) W which, when applied to the observations, recovers estimated sources, $\mathbf{a}(t)$; thus

$$\mathbf{a}(t) = W\mathbf{x}(t).$$

Optimising W to maximise the statistical independence between the components of $\mathbf{a}(t)$ finds estimated sources which are shown in the bottom row of figure 1.1. It is clear that the algorithm has done a good job in separating the sources: the noisy blues recording is estimated together with its noise (plots (a) and (g)), while the string quartet is uncontaminated (plots (b) and (h)). To the ear the recovered sources are indistinguishable from the originals, and in particular there is no trace of music in the unmixed noise.†

Blind source separation has been a practical possibility since the early work of Herault & Jutten [1986] which was analysed from a statistical point of view in [Comon *et al.*, 1991] and further developed by Jutten & Herault [1991], where the phrase ‘independent component analysis’ first appeared. In a seminal paper Comon [1994] proposed the use of mutual information to measure independence and advanced separation algorithms based on approximations to mutual information.

Work by Linsker [1989, 1992] and Nadal & Parga [1994] on mappings which maximise transmitted information showed that the optimal mappings are those which lead to factorised source probability density functions (p.d.f.s). Bell & Sejnowski [1995] and Roth & Barum [1996] each derived stochastic gradient algorithms to find the optimal mapping, and a similar algorithm was suggested by Cardoso & Laheld [1996].

Generative models and maximum likelihood approaches to ICA were proposed and developed by Gaeta & Lacoume [1990] and Pham *et al.* [1992]. However, MacKay [1996], Pearlmutter & Parra [1996] and Cardoso [1997] established that the infomax objective function of Bell & Sejnowski was indeed a likelihood (in the zero noise limit).

Since the mid-nineties there has been an explosion of work on ICA and BSS. Maximum likelihood methods have been extended to incorporate observational noise [Attias, 1999a] and schemes have been developed

† Files with the sources, mixtures and estimated sources may be retrieved from `http://www.dcs.ex.ac.uk/ica`

to permit the separation of sub-Gaussian as well as super-Gaussian† sources (see, for example, [Pham, 1996, Lee *et al.*, 1999b, Everson & Roberts, 1999a]). Pearlmutter & Parra [1996] exploited the temporal structure of sources to improve the separation of timeseries data; extensions of this work appear in Chapter 12 of the present book. Girolami & Fyfe [1997a, 1997b] elucidated the connection between projection pursuit and non-Gaussian sources, and have applied ICA to data mining problems; in Chapter 10 of the present book Girolami gives details of recent work on data classification and visualisation. ICA for non-linear mappings was considered along with early work on linear ICA [Karhunen & Joutsensalo, 1994]. Karhunen describes recent advances in nonlinear ICA in Chapter 4. The generative model formulation of ICA permits Bayesian methods for incorporating prior knowledge, assessing the number of sources and evaluating errors. Early work was done on Bayesian approaches by Roberts [1998] and Knuth [1998a] and more recently by Mohammad-Djafari [1999]. The application of ensemble learning (or variational) methods has greatly simplified the computation required for Bayesian estimates; see Chapter 8 of the present book and [Lappalainen, 1999]. Recent theoretical work (dealt with in the present book) has also examined non-stationary sources (Chapters 5 and 6) and non-stationary mixing (Chapter 11).

Chapter overview This book concentrates mainly on the generative model formulation of ICA as it permits principled extensions. In this introductory chapter we examine ICA from a number of perspectives. Starting from a fairly general point of view, noisy and noiseless models for mixing and the hierarchy of ICA models are discussed first. In subsection 1.2.2 we discuss mutual information as a measure of independence, after which the more general framework of 'contrast functions' is introduced. The introduction of generative models permits maximum likelihood separating matrices to be found; the advantages of a Bayesian approach to ICA are discussed in subsection 1.2.5. ICA has strong links with principal component analysis. PCA and related methodologies are obtained if the sources are Gaussian distributed, as is discussed in section 1.3.

Abandoning Gaussian source distributions permits richer notions of independence to be employed, but also complicates learning the separat-

† A random variable is called sub-Gaussian if its kurtosis is negative and super-Gaussian if its kurtosis is positive. Loosely, the tails of a super-Gaussian p.d.f. decay more slowly than a Gaussian density, while the tails of a sub-Gaussian density decay more rapidly than a Gaussian. See pages 27 and 76.

ing matrix, which can no longer be achieved purely by linear algebra. We attempt to distinguish between the ICA objective or contrast function which is to be extremised and the precise optimisation algorithm. This and the relations between various approaches to noiseless ICA are the subjects of sections 1.4 and 1.6.

Extensions to the basic ICA model are introduced in section 1.8, and finally we briefly describe some applications of ICA.

1.2 Linear mixing

We begin by considering a general model of mixing, which will subsequently be simplified and approximated to permit tractable calculations to be made. The basic model is a discrete time model in which M sources $s_m(t)$ are instantaneously mixed and the resulting mixture, possibly corrupted by noise, is observed. Writing the source signals at the instant t† in vector form, $\mathbf{s}(t) = [s_1(t), s_2(t), \dots, s_M(t)]^\mathsf{T}$, the N-dimensional observations, $\mathbf{x}(t) = [x_1(t), x_2(t), \dots, x_N(t)]^\mathsf{T}$, are generated by a, possibly nonlinear, mixture corrupted by additive observational or sensor noise $\mathbf{n}(t)$ as follows:

$$\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t)) + \mathbf{n}(t), \tag{1.2}$$

where $\mathbf{f} : \mathbb{R}^M \to \mathbb{R}^N$ is an unknown function.

The goal of blind source separation is to invert the mixing function $\mathbf{f}$ and recover the sources. The qualifier *blind* signifies that little is known about the quantities on the right hand side of equation (1.2); the mixing function and the noise and, of course, the sources themselves are unknown and must be estimated. Even with infinite data the unmixing problem is very ill-posed without some additional *a priori* knowledge or assumptions about the sources $\mathbf{s}$, the nature of the mixing $\mathbf{f}$ and the observational noise $\mathbf{n}$. In Chapter 4 Karhunen examines recent approaches to blind source separation with nonlinear mixing. Traditional treatments of ICA, however, make the assumption that the sources are *linearly* mixed by a *mixing matrix* $A \in \mathbb{R}^{N \times M}$. Thus observations are assumed to be generated by

$$\mathbf{x}(t) = A\mathbf{s}(t) + \mathbf{n}(t). \tag{1.3}$$

† Although we call t ‘time’, for most ICA models t is really an index. Most models do not assume any causal dependence of $s_m(t_2)$ on $s_m(t_1)$ when $t_2 > t_1$. See section 1.5 and Chapters 12 and 11.

For simplicity it is usually assumed that **s** and **n** have mean zero, and consequently **x** has zero mean.

Although the nonlinear mixing function has been replaced with an (unknown) matrix the problem of identifying **s** is still under-determined, because there are $N + M$ unknown signals (the noises and the sources) and N known signals (the observations). Progress is only possible with additional assumptions about the nature of the sources and noise.

The principal assumption which permits progress is that the sources are *independent*, which incorporates the idea that each source signal is generated by a process unrelated to the other sources. For example, the voices at a cocktail party might be regarded as independent. Independent Component Analysis is therefore a method for blind source separation, and if independent components can be found they are identified with the (hidden) sources.

1.2.1 Hierarchy of ICA models

Although all ICA models assume the sources to be independent, assumptions about the characteristics of the noise and the source densities lead to a range of ICA models, whose relationships are summarised in figure 1.2.

An important class of models is obtained by assuming that both the sources and noise are Gaussian distributed. Factor Analysis describes the linear model with Gaussian sources and a diagonal noise covariance matrix; restricting the covariance matrix to be isotropic yields Probabilistic Principal Component Analysis (PPCA), and Principal Component Analysis emerges in the absence of noise. These models are described in section 1.3.

Gaussian source models, although historically important and computationally attractive, are, however, seriously limited in their ability to separate sources and recent work on source separation depends crucially on the assumption that the sources are non-Gaussian.

Attias [1999a] has developed an ICA model with linear mixing and observational noise; see Chapter 3 of the present book. The majority of classical ICA models, however, are noiseless so that observations are generated according to

$$\mathbf{x} = A\mathbf{s}. \tag{1.4}$$

Variations of these models depend upon the probabilistic model used for the sources: flexible source models, which depend continuously upon

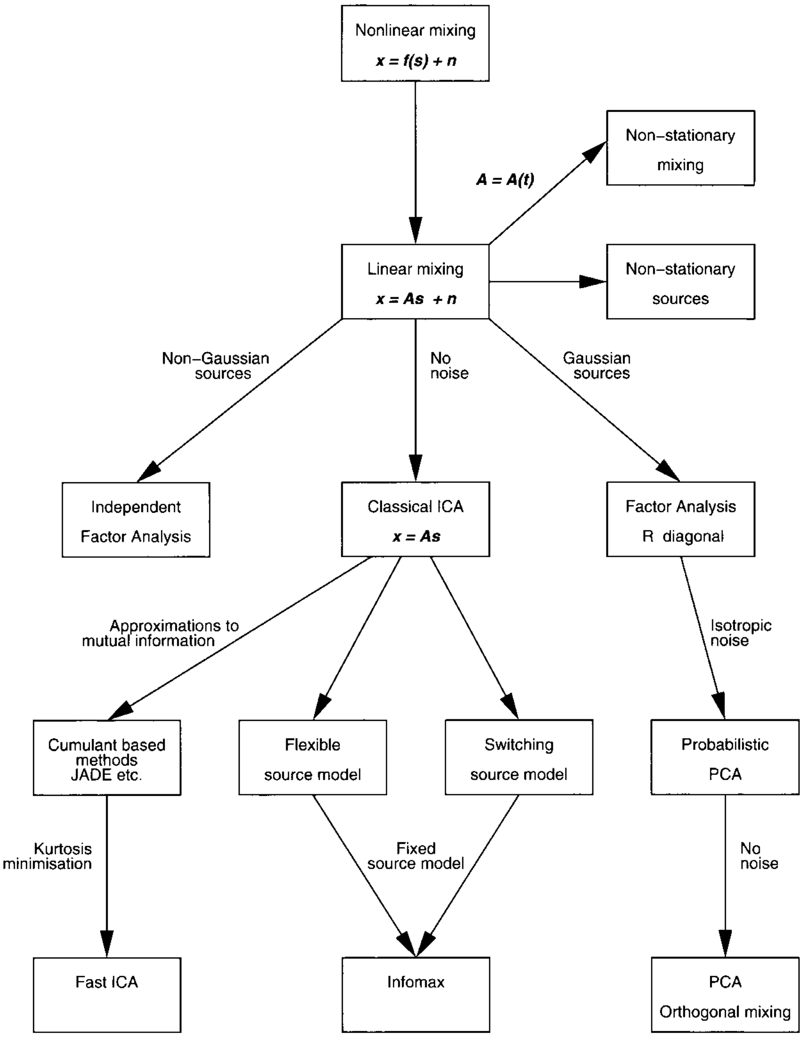

Figure 1.2. ***Hierarchy of ICA Models***

their parameters, and schemes which switch between two source models dependent upon the moments of the recovered sources are discussed in section 1.4. If the source model is fixed to be a single function with no explicit parameters, the Bell & Sejnowski Infomax algorithm [Bell & Sejnowski, 1995] is recovered (subsection 1.5). These models all recover sources which are maximally independent. The degree of independence is measured by the mutual information (subsection 1.2.2) between the recovered sources. Independence between the recovered sources may be approximated by cumulant based expansions. Cumulant based methods are briefly described in section 1.4. An elegant and fast fixed-point technique, FastICA, which maximises the kurtosis of the recovered sources is described by Hyvärinen in Chapter 2.

1.2.2 Independent sources

The assumption underlying all ICA models is that the sources are *independent*. The M sources together generate an M-dimensional probability density function (p.d.f.) $p(\mathbf{s})$. Statistical independence between the sources means that the joint source density factorises as

$$p(\mathbf{s}) = \prod_{m=1}^{M} p(s_m(t)). \tag{1.5}$$

We denote by $\mathbf{a}(t) = [a_1(t), a_2(t), \ldots, a_M(t)]^\mathsf{T}$ the estimates of the true sources $\mathbf{s}(t)$ that are recovered by blind source separation. If the p.d.f. of the estimated sources also factorises then the recovered sources are independent and the separation has been successful. Independence between the recovered sources is measured by their mutual information, which is defined in terms of entropies.

The (differential) entropy of an M-dimensional random variable $\mathbf{x}$ with p.d.f. $p(\mathbf{x})$ is

$$H[\mathbf{x}] = H[p(\mathbf{x})] \overset{\text{def}}{=} -\int p(\mathbf{x}) \log p(\mathbf{x})\, d\mathbf{x}. \tag{1.6}$$

(Square brackets are used to emphasise that the entropy is a statistical quantity that depends on the p.d.f. of $\mathbf{x}$, rather than directly on $\mathbf{x}$ itself.) The entropy measures the average amount of information that $\mathbf{x}$ encodes, or, alternatively, the average amount of information that observation of $\mathbf{x}$ yields [Cover & Thomas, 1991]. If base 2 logarithms are used the entropy is measured in bits.

The joint entropy $H[\mathbf{x}, \mathbf{y}]$ of two random variables $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$H[\mathbf{x}, \mathbf{y}] = -\int p(\mathbf{x}, \mathbf{y}) \log p(\mathbf{x}, \mathbf{y})\, d\mathbf{x}\, d\mathbf{y}. \tag{1.7}$$

The conditional entropy of $\mathbf{x}$ given $\mathbf{y}$ is

$$H[\mathbf{x}\,|\,\mathbf{y}] = -\int p(\mathbf{x}, \mathbf{y}) \log p(\mathbf{x}\,|\,\mathbf{y})\, d\mathbf{x}\, d\mathbf{y}. \tag{1.8}$$

from which it follows that

$$\begin{aligned} H[\mathbf{x}, \mathbf{y}] &= H[\mathbf{x}] + H[\mathbf{y}\,|\,\mathbf{x}] && (1.9)\\ &= H[\mathbf{y}] + H[\mathbf{x}\,|\,\mathbf{y}]. && (1.10) \end{aligned}$$

Equation (1.9) may be interpreted to mean that the (average) information that $\mathbf{x}$ and $\mathbf{y}$ jointly encode is the sum of the information encoded by $\mathbf{x}$ alone and the information encoded by $\mathbf{y}$ given a knowledge of $\mathbf{x}$.

The mutual information between two random variates $\mathbf{x}$ and $\mathbf{y}$ is defined in terms of their entropies.

$$\begin{aligned} I[\mathbf{x}; \mathbf{y}] &\overset{\text{def}}{=} H[\mathbf{x}] + H[\mathbf{y}] - H[\mathbf{x}, \mathbf{y}] && (1.11)\\ &= H[\mathbf{x}] - H[\mathbf{x}\,|\,\mathbf{y}] && (1.12)\\ &= H[\mathbf{y}] - H[\mathbf{y}\,|\,\mathbf{x}]. && (1.13) \end{aligned}$$

The mutual information is thus the difference in the information that is obtained by observing $\mathbf{x}$ and $\mathbf{y}$ separately or jointly. Alternatively, as (1.13) shows, the information $H[\mathbf{x}]$ encoded by $\mathbf{x}$ that cannot be obtained by observing $\mathbf{y}$ is $I[\mathbf{x}; \mathbf{y}]$. The mutual information is zero if and only if $\mathbf{x}$ and $\mathbf{y}$ are independent (i.e., $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$). The mutual information is non-negative [Cover & Thomas, 1991] which follows from the fact that more information may be obtained by observing $\mathbf{x}$ and $\mathbf{y}$ separately than jointly.

With a slight abuse of notation, the mutual information between the *components* of $\mathbf{a}$ (sometimes called the *redundancy* of $\mathbf{a}$) is written as

$$\begin{aligned} I[\mathbf{a}] &\overset{\text{def}}{=} I[\mathbf{a}; \{a_m\}] && (1.14)\\ &= \sum_{m=1}^{M} H[a_m] - H[\mathbf{a}] && (1.15)\\ &= \int p(\mathbf{a}) \log \frac{p(\mathbf{a})}{\prod_{m=1}^{M} p_m(a_m)}\, d\mathbf{a}. && (1.16) \end{aligned}$$

The first term of (1.15) is the sum of the information carried by the recovered sources individually, and $H[\mathbf{a}]$ is the information carried jointly. $I[\mathbf{a}]$ is therefore the information common to the variables and thus measures their independence. It is again non-negative and equal to zero if and only if the components of $\mathbf{a}$ are mutually independent, so that there is no common information and the joint density factorises: $p(\mathbf{a}) =$

$\prod_{m=1}^{M} p(a_m)$. If the estimated sources carry no common information then nothing can be inferred about a recovered source from a knowledge of the others and the recovered sources are independent, $I[\mathbf{a}] = 0$. In this case the blind source separation has been successful.

The Kullback-Leibler (KL) divergence between two p.d.f.s $p(\mathbf{x})$ and $q(\mathbf{x})$ is defined as

$$\mathrm{KL}[p \parallel q] \overset{\mathrm{def}}{=} \int_{\mathbf{x}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})}\, d\mathbf{x}. \tag{1.17}$$

Note that $\mathrm{KL}[p \parallel q] \neq \mathrm{KL}[q \parallel p]$. Comparison of equations (1.16) and (1.17) shows that the mutual information between the recovered sources is identical to the Kullback-Leibler divergence between the joint density $p(\mathbf{a})$ and the factorised density $\prod_{m=1}^{M} p(a_m)$. Independent component analysis attempts therefore to find a separating transform (a matrix when the mixing is linear) that minimises this KL divergence.

Scaling and permutation ambiguities The linear generative model (1.3) introduces a fundamental ambiguity in the scale of the recovered sources. The ambiguity arises because scaling a source by a factor λ ($s_m(t) \mapsto \lambda s_m(t)$) is exactly compensated by dividing the corresponding column of the mixing matrix by λ. In terms of the mutual information, we see that mutual information is independent of the scale of the recovered sources: the degree of independence between variables does not depend upon the units in which they are measured.† Therefore $I[\mathbf{a}] = I[D\mathbf{a}]$ for any diagonal matrix D ($D_{ii} \neq 0$). Furthermore, the order in which the components of $\mathbf{a}$ are listed is immaterial to their independence, so $I[\mathbf{a}] = I[P\mathbf{a}]$ for any permutation matrix P. Putting these together, $I[\mathbf{a}] = I[PD\mathbf{a}]$ which shows that the sources can only be recovered up to an arbitrary permutation and scaling.

In the zero noise limit a separating matrix W, which inverts the mixing, is sought so that $\mathbf{a} = W\mathbf{x}$. In this case, rather than $WA = I$, the best that may be achieved is

$$WA = PD, \tag{1.18}$$

because $I[\mathbf{s}] = I[W\mathbf{s}] = I[PDW\mathbf{s}]$. In the presence of isotropic observational noise the scaling and permutation ambiguities remain. Anisotropic noise destroys the permutation ambiguity, though the scaling ambiguity remains.

† More generally, mutual information is invariant under component-wise invertible transformations [Cover & Thomas, 1991].