A Constrained EM Algorithm for Independent Component Analysis

Max Welling *Markus Weber*

California Institute of Technology 136-93
Pasadena, CA 91125
{welling,rmw}@vision.caltech.edu

Keywords: Independent Component Analysis, Expectation Maximization

Abstract

We introduce a novel way of performing independent component analysis using a constrained version of the expectation maximization algorithm. The source distributions are modeled as D one-dimensional mixtures of Gaussians. The observed data are modeled as linear mixtures of the sources with additive, isotropic noise. This generative model is fit to the data using constrained EM. The simpler "soft-switching" approach is introduced, which uses only one parameter to decide on the sub- or super-Gaussian nature of the sources. It is explained how our approach relates to independent factor analysis.

1 Introduction

Independent component analysis (ICA) addresses the problem of finding the factors that contribute independently (in a statistical sense) to some observed data from a set of sensors. The term ICA is strongly connected to the analysis of linear mixtures and it is usually assumed that the data are noise free. In this sense, ICA can be seen as a generalization of principal component analysis (PCA), which *decorrelates* the data, using only second order statistical information. ICA aims at identifying those directions that are independent across all statistical orders. However, in practical situations this is often impossible, in which case the minimization of different contrast functions can be used to emphasize different aspects of the data.

The problem was first addressed by Jutten and Herault (Jutten and Herault, 1991). Other important landmarks were established by Amari et al. (Amari et al., 1996), Cardoso and Laheld (Cardoso and Laheld, 1996), Comon (Comon, 1994), Hyvärinen (Hyvärinen, 1997), Girolami and Fyfe (Girolami and Fyfe, 1997), Oja (Oja, 1997), Pearlmutter and Parra (Pearlmutter and Parra, 1996) and Bell and Sejnowski (Bell and Sejnowski, 1995). This list is by no means exhaustive.

Taking a biological perspective, it was pointed out by Barlow (Barlow, 1989) that the brain might follow the strategy of recoding information to a less redundant representation. For super-Gaussian sources, this has the effect of producing sparse codes. Results published in (Bell and Sejnowski, 1997) show that the independent components of natural images are edge filters which supports the view that the receptive fields of the visual system do indeed transform the input to a sparser representation.

Applications of ICA include denoising, data visualization (projection pursuit), blind source separation, deconvolution, feature extraction, data compression and pattern recognition. These applications can be found in the fields of radar and sonar signal processing, image processing, seismic signal processing, speech recognition, telecommunication and medical signal processing. Particularly in the last category, ICA has been employed to find independent factors contributing to fMRI and EEG data (McKeown et al., 1998), (Makeig et al., 1997).

In this paper, we derive an EM algorithm that fits a linear model with non-Gaussian sources to the data. The source densities are modeled by one-dimensional Gaussian mixtures, whose means, variances and mixture coefficients are estimated along with the mixing matrix. In addition, we estimate the variance of isotropic noise, superimposed on the data. A simpler "softswitch" version is proposed for situations with fewer available samples.

We are not the first to propose EM for ICA. Noisy ICA was discussed in (Lewicki and Sejnowski, 1998), while the EM algorithm was applied to this problem by (Moulines et al., 1997) and (Attias, 1999). Although the latter approaches allow the estimation of general (Gaussian) noise covariances and non-square mixing matrices, their complexity grows exponentially with the number of sources. Attias does discuss interesting variational approximations to overcome this problem, however, even these are computationally expensive since the E-step has to be solved iteratively. The method we present in this paper is the first for square ICA with isotropic noise, that is both exact and tractable when the number of sources is large.

2 Independent Component Analysis

Let $s_i, \quad i = 1, ..., D$, denote a random variable for source i, and suppose every source is distributed according to a distribution $p_i(s_i)$. Assume that we observe data $x_j, \quad j = 1, ..., K$, from K sensors which are (approximately) *linear* mixtures of the source signals. We also assume that the sensor data are contaminated with uncorrelated Gaussian noise, n_i. In vector notation we thus find the following simple relation between the random variables,

$$\mathbf{x} = A\ \mathbf{s} + \mathbf{n}, \tag{1}$$

where A is the mixing matrix. The task of an ICA algorithm is as follows:

Given a sequence of N data vectors $\mathbf{x}_n, \quad n = 1, ..., N$, retrieve the mixing matrix A and the original source sequence $\mathbf{s}_n, \quad n = 1, ..., N$.

It turns out that this is only possible up to a permutation and a scaling of the original source data.

It has recently been realized that the estimation of the unmixing matrix, $W = A^{-1}$, can be understood in the framework of maximum likelihood estimation (Pearlmutter and Parra, 1996) by fitting the following model density to the data $\mathbf{x}$,

$$p(\mathbf{x}) = p_1(\mathbf{w}_1^T\mathbf{x})\ p_2(\mathbf{w}_2^T\mathbf{x})\ ...\ p_D(\mathbf{w}_D^T\mathbf{x})\ \det(W), \tag{2}$$

where $\det(W)$ is the Jacobian in the transformation rule for densities. This implies that the coordinates $\mathbf{u} = W\mathbf{x}$, where the rows of W are the vectors $\mathbf{w}_1, ..., \mathbf{w}_D$, are independently distributed according to the pdfs $p_1(u_1), ..., p_D(u_D)$. Maximizing the log-likelihood,

$$L(W|\{\mathbf{x}_n\}) = \log[\det(W)] + \frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{D} \log[p_i(\mathbf{w}_i^T \mathbf{x}_n)], \tag{3}$$

with respect to W provides us with the ML estimate for W. We will use EM to find the ML estimate of the mixing matrix A, along with the overall noise variance and the parameters governing the source densities. The sources, $\mathbf{s}$, are therefore treated as "hidden" variables.

3 The Model

The data are modeled by the following generative model. The causes (sources) are represented by a D-dimensional vector $\mathbf{s}$ and are distributed according to the probability density model,

$$p(\mathbf{s}) = \prod_{i=1}^{D} \sum_{z_i=1}^{M_i} p(s_i|z_i)\ p(z_i) = \prod_{i=1}^{D} \sum_{z_i=1}^{M_i} \mathcal{G}_{s_i}[\mu_{z_i}, \sigma_{z_i}^2]\ \pi_{z_i}. \tag{4}$$

Clearly, the assumption is being made that the causes are independently distributed. Every independent direction can have a different density model, which, in anticipation of the EM algorithm, is approximated by a mixture of Gaussians. Here we introduced the D-dimensional vector $\mathbf{z}$, every entry of which can assume an integer value from 1 to M_i, such that its value labels a component of the Gaussian mixture.[1] The term $\mathcal{G}_x[\mu, \sigma^2]$ represents a Gaussian density over x with mean μ and variance σ^2. Without loss of generality, we will assume that the densities $p_i(s_i)$ have unit variance. Any scale factor can be absorbed by the diagonal elements of the mixing matrix A. We assume the sources $\mathbf{s}$ to be mixed linearly under the addition of independent and isotropic Gaussian noise with zero mean and variance β^2,

$$\mathbf{x} = A\ \mathbf{s} + \mathbf{n}, \qquad \mathbf{n} \sim \mathcal{G}_{\mathbf{n}}[0, \beta^2 I]. \tag{5}$$

In the EM algorithm (see next section) we need to calculate the *posterior* density $p(\mathbf{s}, \mathbf{z}|\mathbf{x})$. Using Bayes rule,

$$p(\mathbf{s}, \mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{s})\ p(\mathbf{s}|\mathbf{z})\ p(\mathbf{z})}{p(\mathbf{x})}, \tag{6}$$

we observe that we need to calculate $p(\mathbf{x})$ as well. It is given by,

$$p(\mathbf{x}) = \int d\mathbf{s}\ p(\mathbf{x}|\mathbf{s})\ p(\mathbf{s}) = \int d\mathbf{s}\ \mathcal{G}_{\mathbf{x}}[A\mathbf{s}, \beta^2 I]\ \prod_{i=1}^{D} \sum_{z_i=1}^{M_i} \pi_{z_i}\ \mathcal{G}_{s_i}[\mu_{z_i}, \sigma_{z_i}^2]. \tag{7}$$

To compute the integral over $\mathbf{s}$ we can rewrite,

$$\mathcal{G}_{\mathbf{x}}[A\mathbf{s}, \beta^2 I] = \frac{1}{\sqrt{\det(A^T A)}}\ \mathcal{G}_{\mathbf{s}}[\mathbf{u}, \beta^2 (A^T A)^{-1}], \qquad \mathbf{u} = A^{-1}\mathbf{x}. \tag{8}$$

[1] Notice that the explicit components of, say, π_{z_i} are given by $\pi_{z_i}^i$. In the following we will slightly abuse notation and ignore the upper index i since it is already present in the lower index z_i.

The integral is complicated due to the fact that the Gaussian cannot be factored because of its non-diagonal covariance. This can be remedied by sphering the data in a preprocessing step. The subsequent ICA algorithm will reduce all higher order dependencies under the constraint that the data stay decorrelated at all times. Consequently, the random variables $\mathbf{x}$ have covariance equal to the identity matrix. Using this and the invertibility of A, we obtain the following condition on the mixing matrix,

$$A\ \mathbf{E}[\mathbf{s}\mathbf{s}^T]\ A^T + \mathbf{E}[\mathbf{n}\mathbf{n}^T] = \mathbf{E}[\mathbf{x}\mathbf{x}^T] \Rightarrow AA^T = A^TA = (1-\beta^2)I. \tag{9}$$

where $\mathbf{E}[.]$ denotes expectation. According to (9) we may reparametrize A as a scaling times a rotation,

$$A = \sqrt{1-\beta^2}\ R \qquad RR^T = R^TR = I. \tag{10}$$

Now R encodes all relevant information, since the sources can only be retrieved up to an arbitrary scale. The signal to noise ratio is reflected by the scale-factor.

Finally, using (4), (6), we find for the posterior,

$$p(\mathbf{s},\mathbf{z}|\mathbf{x}) = \prod_{i=1}^{D} \alpha_{z_i}\ \mathcal{G}_{s_i}[b_{z_i}, \gamma_{z_i}^2], \tag{11}$$

where

$$\gamma_{z_i}^2 = \frac{\beta^2\ \sigma_{z_i}^2}{(1-\beta^2)\sigma_{z_i}^2 + \beta^2}, \qquad b_{z_i} = \gamma_{z_i}^2 \left(\frac{1-\beta^2}{\beta^2} u_i + \frac{\mu_{z_i}}{\sigma_{z_i}^2} \right), \qquad \alpha_{z_i} = \frac{\pi_{z_i}\ \mathcal{G}_{u_i}[\mu_{z_i}, \frac{\beta^2}{1-\beta^2} + \sigma_{z_i}^2]}{\sum_{w_i=1}^{M_i} \pi_{w_i}\ \mathcal{G}_{u_i}[\mu_{w_i}, \frac{\beta^2}{1-\beta^2} + \sigma_{w_i}^2]}. \tag{12}$$

The fact that the posterior factorizes enables us to treat the E-step as D one-dimensional subproblems. This, in turn, makes the case of many sources tractable. From (11) we easily derive the following densities, needed in the next section,

$$p(\mathbf{s}|\mathbf{x}) = \prod_{i=1}^{D} \sum_{z_i=1}^{M_i} \alpha_{z_i}\ G_{s_i}[b_{z_i}, \gamma_{z_i}^2], \qquad p(z_i|\mathbf{x}) = \alpha_{z_i}, \qquad p(s_i|z_i,\mathbf{x}) = \mathcal{G}_{s_i}[b_{z_i}, \gamma_{z_i}^2]. \tag{13}$$

4 The Expectation Maximization Algorithm

4.1 The M-step

Suppose we have N samples, identically and independently drawn from the distribution $p(\mathbf{x})$ in (7). Independent component analysis may be interpreted as finding the matrix, R, that best fits the generative model to these data. Due to the orthogonality of R this implies the estimation of $\frac{1}{2}D(D-1)$ parameters on a constraint manifold. Simultaneously, we estimate the parameters of the source distributions μ_{z_i}, σ_{z_i}, and π_{z_i}, and the noise parameter β. The set of all parameters is denoted by θ.

Following standard EM procedure we will maximize the function,

$$Q(\tilde{\theta}|\theta) = \frac{1}{N}\sum_{n=1}^{N} \int d\mathbf{s} \sum_{\mathbf{z}}\ p(\mathbf{s},\mathbf{z}|\mathbf{x}_n;\theta)\ \log\{p(\mathbf{x}_n|\mathbf{s};\tilde{\theta})\ p(\mathbf{s}|\mathbf{z};\tilde{\theta})\ p(\mathbf{z}|\tilde{\theta})\}, \tag{14}$$

where $\tilde{\theta}$ denotes the new parameters with respect to which we optimize $Q(\tilde{\theta}|\theta)$, and θ are the old values from the previous iteration that we keep fixed in the M-step. The expression (14) can be broken down into a sum of three terms. The first term,

$$Q_1(\tilde{R},\tilde{\beta}|R,\beta) = \frac{1}{N}\sum_{n=1}^{N}\int d\mathbf{s}\; p(\mathbf{s}|\mathbf{x}_n)\;\log\{p(\mathbf{x}_n|\mathbf{s})\}, \tag{15}$$

is the part that needs to be maximized with respect to $\tilde{R}$ and $\tilde{\beta}$, under the constraint $\tilde{R}^T\tilde{R} = I$, while $p(\mathbf{s}|\mathbf{x}_n)$ is given by (13). To enforce the constraint we will introduce a symmetric matrix of Lagrange multipliers Λ and add a constraint term, $\mathbf{tr}[\Lambda(\tilde{R}^T\tilde{R}-I)]$, to Q_1. Differentiating Q_1 with respect to both $\tilde{R}$ and Λ and equating to zero yields the following update rule for the rotation $\tilde{R}$,

$$\tilde{R} = V\;(V^TV)^{-\frac{1}{2}}, \qquad V \equiv \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\langle\mathbf{s}^T\rangle_n \tag{16}$$

where $\langle . \rangle_n$ is defined as the expectation over $p(\mathbf{s}|\mathbf{x}_n)$, and taking a fractional power is defined by performing an eigenvalue decomposition and taking the fractional power of the eigenvalues.

Introducing the scale-factor $\tilde{\omega} = \sqrt{(1-\tilde{\beta}^2)}$, differentiating Q_1 with respect to $\tilde{\omega}$ and equating to zero leaves us with the following cubic equation,

$$\tilde{\omega}^3 - a_1\;\tilde{\omega}^2 + a_2\;\tilde{\omega} - a_1 = 0, \tag{17}$$

where

$$a_1 = \frac{1}{N\;D}\sum_{n=1}^{N}\mathbf{tr}\left[\tilde{R}\;\langle\mathbf{s}\rangle_n\mathbf{x}_n^T\right], \qquad a_2 = \frac{1}{N\;D}\sum_{n=1}^{N}\mathbf{tr}\left[\langle\mathbf{s}\mathbf{s}^T\rangle_n\right] \tag{18}$$

Analytic expressions for roots of a cubic equation exist (see for instance (Press et al., 1988)).

For both the second and the third term of (14) we may observe that they can be written as a sum over D independent terms,

$$Q_2(\tilde{\mu},\tilde{\sigma}|\mu,\sigma) \;=\; \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{D}\int ds_i\sum_{z_i=1}^{M_i}p(s_i,z_i|\mathbf{x}_n)\log p(s_i|z_i) \tag{19}$$

$$Q_3(\tilde{\pi}|\pi) \;=\; \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{D}\sum_{z_i=1}^{M_i}p(z_i|\mathbf{x}_n)\log p(z_i). \tag{20}$$

Using (13) we can derive the following update rules for $\tilde{\pi}$, $\tilde{\mu}$ and $\tilde{\sigma}$,

$$\tilde{\pi}_{z_i} = \frac{1}{N}\sum_{n=1}^{N}\alpha_{z_i}^n, \qquad \tilde{\mu}_{z_i} = \frac{\sum_{n=1}^{N}\alpha_{z_i}^n\langle s_i\rangle_{z_i}^n}{\sum_{n=1}^{N}\alpha_{z_i}^n}, \qquad \tilde{\sigma}_{z_i}^2 = \frac{\sum_{n=1}^{N}\alpha_{z_i}^n\langle s_i^2\rangle_{z_i}^n}{\sum_{n=1}^{N}\alpha_{z_i}^n} - \tilde{\mu}_{z_i}^2. \tag{21}$$

where $\langle . \rangle_{z_i}^n$ denotes taking the average with respect to $p(s_i|z_i,\mathbf{x}_n)$.

4.2 The E-step

In the E-step we have to calculate the sufficient statistics $\langle s_i \rangle^n_{z_i}$, $\langle s_i^2 \rangle^n_{z_i}$, $\langle s_i \rangle_n$ and $\langle s_i^2 \rangle_n$. Notice that we don't have to calculate the off-diagonal terms of the covariances between the sources given the data, $\langle s_i s_j \rangle$. The former statistics can be computed using (11),(12) and (13):

$$\langle s_i \rangle^n_{z_i} = b^n_{z_i}, \qquad \langle s_i \rangle_n = \sum_{z_i=1}^{M_i} \alpha^n_{z_i}\ b^n_{z_i} \tag{22}$$

$$\langle s_i^2 \rangle^n_{z_i} = \gamma^2_{z_i} + (b^n_{z_i})^2, \qquad \langle s_i^2 \rangle_n = \sum_{z_i=1}^{M_i} \alpha^n_{z_i}\ (\gamma^2_{z_i} + (b^n_{z_i})^2) \tag{23}$$

4.3 Soft-Switching

The algorithm derived in the previous sections adaptively estimates the source densities through the estimation of $3 \times M_i$ parameters for source i, where M_i is the number of Gaussian components in the mixture. However, it is well known that the separation result is relatively insensitive to the precise shape of the source densities, as long as the sign of the kurtosis is correct. The question naturally arises, if we can use simpler source models. The method advocated in the *Extended Infomax Algorithm* (Lee et al., 1999) is to switch between a sub- and a super-Gaussian model depending on the sign of a switching moment, which tests the stability of the solution (Cardoso and Laheld, 1996). This method only requires the estimation of one binary parameter per source. Trying to incorporate this idea in our EM framework, we found that in the presence of (near) Gaussian sources the algorithm oscillates and fails to converge. This property can be understood by observing that a Gaussian source has zero kurtosis. Therefore, the algorithm cannot decide whether to use a super-Gaussian or a sub-Gaussian mixture. The same behaviour was found in the *Extended Infomax Algorithm* but causes less severe problems because only small steps are taken at every iteration. The next simplest solution involves a "soft" interpolation between one model and the other by assuming a certain probability, r, that the source is super-Gaussian and a probability $1 - r$ that it is sub-Gaussian. EM can then easily find the best estimate for r. We split the Gaussian component densities for every source into a set that is able to model a super-Gaussian density and a set that is able to model a sub-Gaussian density, i.e.

$$\pi_{z_i} = \begin{cases} r_i\ \rho_{z_i} & z_i \in [1, ..., K_i] \\ (1 - r_i)\ \rho_{z_i} & z_i \in [K_i + 1, ..., M_i]. \end{cases} \tag{24}$$

The values of ρ_{z_i} are fixed and chosen such that for every i the mixture coefficients with $z_i = 1, ..., K_i$ describe a super-Gaussian density and the coefficients with $z_i = K_i + 1, ..., M_i$ describe a sub-Gaussian density. Both sets of mixture coefficients sum independently to 1: $\sum_{z_i=1}^{K_i} \rho_{z_i} = 1, \sum_{z_i=K_i+1}^{M_i} \rho_{z_i} = 1$.

The only parameters that are going to be updated are the r_i. The update rule is given by,

$$\tilde{r}_i = \frac{1}{N} \sum_{n=1}^{N} \sum_{z_i=1}^{K_i} \alpha^n_{z_i}. \tag{25}$$

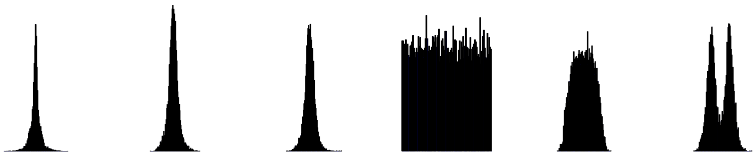

Figure 1: Histograms of data from 3 sound sources (first 3) and 3 artificial sources (last 3).

For every source, the algorithm will estimate the optimal mixture of these two model densities. For super-Gaussian sources the value of r is higher than for sub-Gaussian sources.

4.4 Source Reconstruction

The most straightforward manner to estimate the original source samples is to apply the inverse mixing matrix to the observed data, $\hat{\mathbf{s}}_n = A^{-1}\mathbf{x}_n$. However, if the noise variance β has a significant value this procedure is sub-optimal. A better way is to consider the MAP-estimate, given by the maximum of the posterior density. Because the posterior density factorizes, the computation of the MAP estimate can be done separately for every source,

$$\hat{s}_{in} = \text{arg max}_{s_i} \;\; p(s_i|\mathbf{x}_n) = \text{arg max}_{s_i} \;\; \sum_{z_i=1}^{M_i} \alpha_{z_i}^n \; \mathcal{G}_{s_i}[b_{z_i}^n, \gamma_{z_i}^2], \qquad i = 1...D, \tag{26}$$

where we used (13). The solution to these D one dimensional optimization problems can be found using standard techniques.

5 An Example: Sound Sources and Artificial Data

To test the validity of our algorithm we mixed data from 3 sound sources[2] together with 3 artificial sources. The sound data had a super-Gaussian profile while the artificial data were more uniformly or multimodally distributed (see Figure 1). The CD recordings were subsampled by a factor of 5, resulting in 10^3 samples. The mixing matrix consisted of 1s on the diagonal and 0.25 in the off-diagonal entries. No noise was added to the mixed sources. The algorithm was run in both its fully adaptive mode (two Gaussians per source), as well as in the softswitch mode (a mixture of two Gaussian mixtures consisting of two Gaussians each). As a stopping criterion we chose a threshold for the increase in log-likelihood. In Figure 2 we show the evolution over time of the Amari distance[3] between true and estimated mixing matrix, the log-likelihood, the noise variance and the mixing coefficients (softswitch only). The results for the fully adaptive algorithm (dashed lines) and the softswitch approach (solid lines) are very similar, although convergence was faster in the softswitch mode. Furthermore, the estimated noise variance in the fully

[2]The CD recordings can be found at *http://sweat.cs.unm.edu/bap/demos.html.*

[3]The Amari distance(Amari et al., 1996) measures a distance between two matrices up to permutations and scalings: $\mathcal{N} = \sum_{i=1}^N \left(\sum_{j=1}^N \frac{|p_{ij}|}{\max_k |p_{ik}|} - 1 \right) + \sum_{j=1}^N \left(\sum_{i=1}^N \frac{|p_{ij}|}{\max_k |p_{kj}|} - 1 \right)$

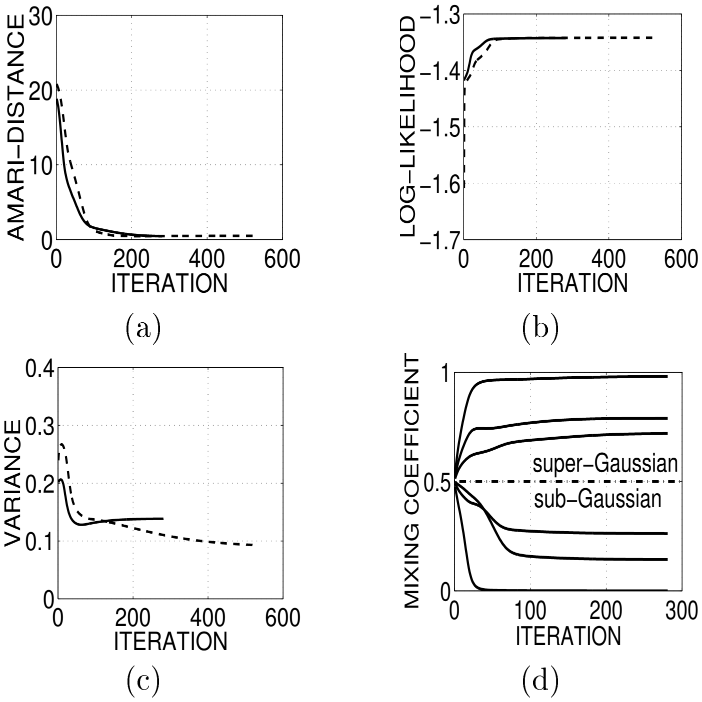

Figure 2: Evolution of Amari distance (a), log-likelihood (b), noise variance (c) and mixing coefficients (d) over time. Solid lines indicate the results for the softswitch algorithm, while dashed lines show the results for the fully adaptive approach.

adaptive mode was lower, due to the greater flexibility in modeling the source densities. From the mixing coefficients (Figure 2d) it can be seen that 3 super-Gaussian (sound) sources and 3 sub-Gaussian (artificial) sources were recovered. The estimated mixing matrices for both approaches, after correcting for scale and permutation, are given below.

$$A_{\mathbf{adaptive}} = \begin{pmatrix} 1 & 0.257 & 0.256 & 0.258 & 0.254 & 0.241 \\ 0.229 & 1 & 0.256 & 0.246 & 0.249 & 0.243 \\ 0.241 & 0.256 & 1 & 0.248 & 0.275 & 0.253 \\ 0.239 & 0.248 & 0.252 & 1 & 0.247 & 0.243 \\ 0.225 & 0.253 & 0.229 & 0.255 & 1 & 0.237 \\ 0.246 & 0.245 & 0.246 & 0.252 & 0.256 & 1 \end{pmatrix}$$

$$A_{\mathbf{softswitch}} = \begin{pmatrix} 1 & 0.255 & 0.265 & 0.261 & 0.245 & 0.241 \\ 0.227 & 1 & 0.273 & 0.255 & 0.235 & 0.244 \\ 0.241 & 0.244 & 1 & 0.251 & 0.239 & 0.247 \\ 0.237 & 0.238 & 0.261 & 1 & 0.238 & 0.243 \\ 0.224 & 0.255 & 0.265 & 0.258 & 1 & 0.239 \\ 0.244 & 0.241 & 0.261 & 0.254 & 0.244 & 1 \end{pmatrix}$$

6 Discussion and Conclusions

This paper proposes to use the EM algorithm to perform ICA. The source vector $\mathbf{s}$ is considered as a hidden variable which is linearly related to an observed vector $\mathbf{x}$. Isotropic noise is added to the observed data vectors. The source distributions are modeled by one-dimensional mixtures of Gaussians. Usually, two to three Gaussians per source are sufficient. The EM algorithm is employed to fit a linear combination of these source distributions to the observed data. The parameters to be estimated are the mixing matrix, the various parameters determining the source densities, as well as the noise variance.

It is well known that the estimation of the mixing matrix is relatively insensitive to the details of the source densities. The most important piece of information is the kurtosis of the source densities. For the case where focus is on source separation (instead of density estimation), a simpler algorithm was developed which uses a mixture of a sub- and a super-Gaussian model for the source densities. One parameter per source, measuring the probability that the data in that direction are explained by the super-Gaussian model (as opposed to being explained by the sub-Gaussian model) is estimated from the data. This "soft-switching" approach avoids oscillations in the presence of Gaussian directions in the data, which have been observed in approaches where a hard decision is made to switch between models.

An important feature of our EM algorithm is that it solves for a mixing matrix in the maximization step which is subject to an orthogonality constraint. This simplifies the algorithm in many ways. Most importantly, it allows to factor the posterior density $p(\mathbf{s}, \mathbf{z}|\mathbf{x})$. This avoids an exponential increase of complexity with the number of sources, and allows an exact treatment in the case of many sources. Furthermore, we only need to calculate two simple sufficient statistics: the source mean and the source variance given the data. Calculation of the source covariances given the data is avoided. The constraint also removes an ambiguity regarding the retrieved scale of the source distributions.

On the other hand, the orthogonality constraint implies that we can only estimate a square mixing matrix, and are restricted to model simple isotropic noise. In some situations it is desirable to explain the data by a lower dimensional space of causes, much in the spirit of factor analysis or principal component analysis. There are at least two ways to avoid this problem: If the interesting directions contain most of the power of the signal, PCA can be used to first reduce the dimensionality of the problem before EM-ICA is employed. Another possibility is simply to run EM-ICA on the complete problem and later reject the Gaussian directions by measuring, e.g. the kurtosis. Too many *Gaussian* sources have the effect of slowing down the convergence and compromising the accuracy of the result.

The idea of using EM has been proposed before in (Attias, 1999) and (Moulines et al., 1997). Our method differs by introducing a constraint on the mixing matrix and the use of an "orthogonal" version of EM. Also, the soft-switch approach, used to encompass different sorts of source distributions, is new to our knowledge.

To understand the difference between EM-ICA and similar, related methods we provide the table in Figure 3(a). It lists several data analysis techniques which linearly relate the sources to the observations. The simplest linear models use Gaussian source distributions (top row) with increasing sophistication in their noise models (from left to right). In probabilistic PCA (Tipping and Bishop, 1997) a simple istropic noise model is used to capture the variance in the directions orthogonal to the principal components. Factor analysis is obtained when more general Gaussian noise is allowed. The bottom row still relates the sources linearly to the observed data but uses a non-Gaussian factorized distribution for these sources. While standard ICA searches for independent projections of the data, EM-ICA fits a probability density, i.e. alongside the estimation of a mixing matrix also the parameters of the source distributions are estimated. This is advantageous in applications like pattern recognition, and opens the possibility of data synthesis. Independent factor analysis (Attias, 1999) is an extension of EM-ICA where more general Gaussian noise is modeled. An important property

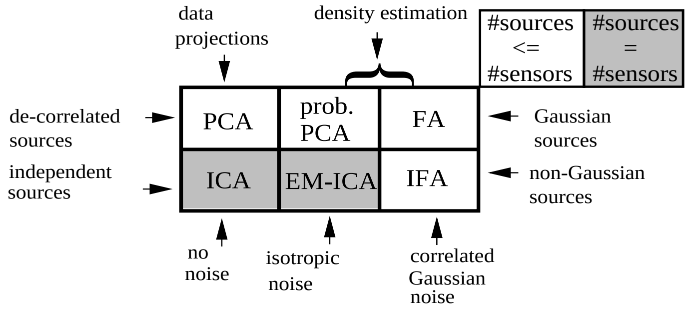

Figure 3: (a) Table of techniques for data analysis. The box contains six methods which relate the observed data linearly to hidden sources. From left to right, top to bottom we have: principle component analysis, probabilistic principle component analysis, factor analysis, independent component analysis, EM-independent component analysis and independent factor analysis. Methods in shaded boxes estimate square mixing matrices, while the other methods allow dimensionality reduction. The properties written around the box apply to the complete row or column of the table.

that IFA shares with PCA, probabilistic PCA and FA, but not with ICA and EM-ICA , is that the number of sources do not have to match the number of sensors. The disadvantage of IFA is, however, that the complexity of the algorithm grows exponentially in the number of sources. The simple noise model and the invertibility of the mixing matrix result in linear complexity for EM-ICA in the number of sources. This implies that, for a large number of sources (10 or more), EM-ICA is still exact while IFA has to rely on approximate variational methods. However, even realistic variational approximations are computationally expensive since the E-phase has to be solved iteratively at each step of the EM algorithm.

The EM framework is ideally suited for treatment of missing data. An extension of EM-ICA , which performs ICA on incomplete data was studied in (Welling and Weber, 1999).

7 Acknowledgements

We would like to thank Pietro Perona for stimulating discussions and the referees for many suggestions that improved the text significantly. M. Welling would like to acknowledge the Sloan Center for its ongoing financial support.

References

Amari, S., Cichocki, A., and Yang, H. (1996). A new algorithm for blind signal separation. *Advances in Neural Information Processing Systems*, 8:757–763.

Attias, H. (1999). Independent factor analysis. *Neural Computation*, 11:803–851.

Barlow, H. (1989). Unsupervised learning. *Neural Computation*, 1:295–311.

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

Bell, A. and Sejnowski, T. (1997). The independent components of natural scenes are edge filters. *Vision Research*, 37:3327–3338.

Cardoso, J. and Laheld, B. (1996). Equivariant adaptive source separation. *IEEE Transactions on signal Processing*, 44:3017–3030.

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

Girolami, M. and Fyfe, C. (1997). An extended exploratory projection pursuit network with linear and nonlinear anti-hebbian lateral connections applied to the cocktail party problem. *Neural Networks*, 10:1607–1618.

Hyvärinen, A. (1997). Independent component analysis by minimization of mutual information. Technical report, Helsinki University of Technology, Laboratory of Computer and Information Science.

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

Lee, T., Girolami, M., and Sejnowski, T. (1999). Independent component analysis using an extended infomax algorithm for mixed sub-gaussian and super-gaussian sources. *Neural Computation*, 11:409–433.

Lewicki, M. and Sejnowski, T. (1998). Learning overcomplete representations. Technical report, Howard Hughes Medical Institute, Computational Neurobiology Lab, The Salk Institute. Accepted for publication in Neural Computation.

Makeig, S., Bell, A., Jung, T., and Sejnowski, T. (1997). Blind separation of auditory event-related brain responses into independent components. *Proceedings of the National Academy of Sciences*, 94:10979–10984.

McKeown, M., Jung, T., Makeig, S., Brown, G., Kinderman, S., Lee, T., and Sejnowski, T. (1998). Spatially independent activity patterns in functional mri data during stroop color-naming task. *Proceedings of the National Academy of Sciences*, 95:803–810.

Moulines, E., Cardoso, J., and Gassiat, E. (1997). Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *Proc. ICASSP*.

Oja, E. (1997). The nonlinear pca learning rule in independent component analysis. *Neural Computation*, 17:25–45.

Pearlmutter, B. and Parra, L. (1996). A context sensitive generalization of ica. *International Conference on Neural Information Processing*, pages 151–157.

Press, W., Flannery, B., Teukolsky, S., and Vetterling, W. (1988). *Numerical recipes in C.* Cambridge university press.

Tipping, M. and Bishop, C. (1997). Probabilistic principal component analysis. Technical Report NCRG/97/010, Aston University, Department of Computer Science and Applied Mathematics, Neural Computing Research Group.

Welling, M. and Weber, M. (1999). A constrained e.m. algorithm for independent component analysis. In *Proceedings of the 6th Joint Symposium on Neural Computation.*