

## A Constrained EM Algorithm for Independent Component Analysis

Max Welling  
Markus Weber

*California Institute of Technology, Pasadena, CA 91125, U.S.A.*

**We introduce a novel way of performing independent component analysis using a constrained version of the expectation-maximization (EM) 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 supergaussian nature of the sources. We explain 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 Héroult (1991). Other important landmarks were established by Amari, Cichocki, and Yang (1996), Cardoso and Laheld (1996), Comon (1994), Hyvärinen (1997), Girolami and Fyfe (1997), Oja (1997), Pearlmutter and Parra (1996), and Bell and Sejnowski (1995). This list is by no means exhaustive.

Taking a biological perspective, Barlow (1989) pointed out that the brain might follow the strategy of recoding information to a less redundant representation. For supergaussian 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 support 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 functional magnetic resonance and electroencephalogram data (McKeown et al., 1998; Makeig, Bell, Jung, & Sejnowski, 1997).

In this article we derive an expectation-maximization (EM) algorithm that fits a linear model with nongaussian 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 “soft-switch” 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 (2000), and the EM algorithm was applied to this problem by Moulines, Cardoso, and Gassiat (1997) and Attias (1999). Although the latter approaches allow the estimation of general (gaussian) noise covariances and nonsquare 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 here 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  $s_i$ ,  $i = 1, \dots, D$ , 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  $x_j$ ,  $j = 1, \dots, K$ , 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,

$$\mathbf{x} = A \mathbf{s} + \mathbf{n}, \quad (2.1)$$

where  $A$  is the mixing matrix. The task of an ICA algorithm is as follows: Given a sequence of  $N$  data vectors  $\mathbf{x}_n$ ,  $n = 1, \dots, N$ , retrieve the mixing matrix  $A$  and the original source sequence  $\mathbf{s}_n$ ,  $n = 1, \dots, N$ . It turns out that this is possible only 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 es-

timization (Pearlmutter & Parra, 1996) by fitting the following model density to the data  $\mathbf{x}$ ,

$$p(\mathbf{x}) = p_1(\mathbf{w}_1^T \mathbf{x}) p_2(\mathbf{w}_2^T \mathbf{x}), \dots, p_D(\mathbf{w}_D^T \mathbf{x}) \det(W), \quad (2.2)$$

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  $\mathbf{w}_1, \dots, \mathbf{w}_D$ , are independently distributed according to the pdfs  $p_1(u_1), \dots, p_D(u_D)$ . Maximizing the log-likelihood,

$$L(W|\{\mathbf{x}_n\}) = \log[\det(W)] + \frac{1}{N} \sum_{n=1}^N \sum_{i=1}^D \log[p_i(\mathbf{w}_i^T \mathbf{x}_n)], \quad (2.3)$$

with respect to  $W$  provides us with the maximum likelihood (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,

$$p(\mathbf{s}) = \prod_{i=1}^D \sum_{z_i=1}^{M_i} p(s_i|z_i) p(z_i) = \prod_{i=1}^D \sum_{z_i=1}^{M_i} \mathcal{G}_{s_i}[\mu_{z_i}, \sigma_{z_i}^2] \pi_{z_i}. \quad (3.1)$$

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.<sup>1</sup> The term  $\mathcal{G}_x[\mu, \sigma^2]$  represents a gaussian density over  $x$  with mean  $\mu$  and variance  $\sigma^2$ . 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  $\beta^2$ ,

$$\mathbf{x} = A \mathbf{s} + \mathbf{n}, \quad \mathbf{n} \sim \mathcal{G}_{\mathbf{n}}[0, \beta^2 I]. \quad (3.2)$$

<sup>1</sup> Notice that the explicit components of, say,  $\pi_{z_i}$  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  $z_i$ .

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

$$p(\mathbf{s}, \mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{s}) p(\mathbf{s}|\mathbf{z}) p(\mathbf{z})}{p(\mathbf{x})}, \quad (3.3)$$

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

$$p(\mathbf{x}) = \int d\mathbf{s} p(\mathbf{x}|\mathbf{s}) p(\mathbf{s}) = \int d\mathbf{s} \mathcal{G}_x[A\mathbf{s}, \beta^2 I] \prod_{i=1}^D \sum_{z_i=1}^{M_i} \pi_{z_i} \mathcal{G}_{s_i}[\mu_{z_i}, \sigma_{z_i}^2]. \quad (3.4)$$

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

$$\mathcal{G}_x[A\mathbf{s}, \beta^2 I] = \frac{1}{\sqrt{\det(A^T A)}} \mathcal{G}_s[\mathbf{u}, \beta^2 (A^T A)^{-1}], \quad \mathbf{u} = A^{-1} \mathbf{x}. \quad (3.5)$$

The integral is complicated due to the fact that the gaussian cannot be factored because of its nondiagonal 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,

$$A \mathbf{E}[\mathbf{s}\mathbf{s}^T] A^T + \mathbf{E}[\mathbf{n}\mathbf{n}^T] = \mathbf{E}[\mathbf{x}\mathbf{x}^T] \Rightarrow AA^T = A^T A = (1 - \beta^2)I, \quad (3.6)$$

where  $\mathbf{E}[\cdot]$  denotes expectation. According to equation 3.6, we may reparameterize  $A$  as a scaling times a rotation,

$$A = \sqrt{1 - \beta^2} R \quad RR^T = R^T R = I. \quad (3.7)$$

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

Finally, using equations 3.1 and 3.3, we find for the posterior,

$$p(\mathbf{s}, \mathbf{z}|\mathbf{x}) = \prod_{i=1}^D \alpha_{z_i} \mathcal{G}_{s_i}[b_{z_i}, \gamma_{z_i}^2], \quad (3.8)$$

where

$$\begin{aligned} \gamma_{z_i}^2 &= \frac{\beta^2 \sigma_{z_i}^2}{(1 - \beta^2)\sigma_{z_i}^2 + \beta^2}, & b_{z_i} &= \gamma_{z_i}^2 \left( \frac{1 - \beta^2}{\beta^2} u_i + \frac{\mu_{z_i}}{\sigma_{z_i}^2} \right), \\ \alpha_{z_i} &= \frac{\pi_{z_i} \mathcal{G}_{u_i}[\mu_{z_i}, \frac{\beta^2}{1 - \beta^2} + \sigma_{z_i}^2]}{\sum_{w_i=1}^{M_i} \pi_{w_i} \mathcal{G}_{u_i}[\mu_{w_i}, \frac{\beta^2}{1 - \beta^2} + \sigma_{w_i}^2]}. \end{aligned} \quad (3.9)$$

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 equation 3.8, we easily derive the following densities, needed in the next section:

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

$$p(z_i|\mathbf{x}) = \alpha_{z_i}, \quad p(s_i|z_i, \mathbf{x}) = \mathcal{G}_{s_i}[b_{z_i}, \gamma_{z_i}^2]. \quad (3.10)$$

#### 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 equation 3.4. ICA 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  $\mu_{z_i}$ ,  $\sigma_{z_i}$ , and  $\pi_{z_i}$  and the noise parameter  $\beta$ . The set of all parameters is denoted by  $\theta$ .

Following standard EM procedure, we will maximize the function,

$$Q(\tilde{\theta}|\theta) = \frac{1}{N} \sum_{n=1}^N \int ds \sum_{\mathbf{z}} p(\mathbf{s}, \mathbf{z}|\mathbf{x}_n; \theta) \log\{p(\mathbf{x}_n|\mathbf{s}; \tilde{\theta}) p(\mathbf{s}|\mathbf{z}; \tilde{\theta}) p(\mathbf{z}|\tilde{\theta})\}, \quad (4.1)$$

where  $\tilde{\theta}$  denotes the new parameters with respect to which we optimize  $Q(\tilde{\theta}|\theta)$ , and  $\theta$  are the old values from the previous iteration that we keep fixed in the M-step. The expression 4.1 can be broken down into a sum of three terms. The first term,

$$Q_1(\tilde{R}, \tilde{\beta}|R, \beta) = \frac{1}{N} \sum_{n=1}^N \int ds p(\mathbf{s}|\mathbf{x}_n) \log\{p(\mathbf{x}_n|\mathbf{s})\}, \quad (4.2)$$

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 equation 3.10. To enforce the constraint, we will introduce a symmetric matrix of Lagrange multipliers  $\Lambda$  and add a constraint term,  $\text{tr}[\Lambda(\tilde{R}^T \tilde{R} - I)]$ , to  $Q_1$ . Differentiating  $Q_1$  with respect to both  $\tilde{R}$  and  $\Lambda$  and equating to zero yields the following update rule for the rotation  $\tilde{R}$ ,

$$\tilde{R} = V (V^T V)^{-\frac{1}{2}}, \quad V \equiv \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n (\mathbf{s}^T)_n, \quad (4.3)$$

where  $(\cdot)_n$  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,

$$\tilde{\omega}^3 - a_1 \tilde{\omega}^2 + a_2 \tilde{\omega} - a_1 = 0, \quad (4.4)$$

where

$$a_1 = \frac{1}{ND} \sum_{n=1}^N \text{tr} \left[ \tilde{R} \langle \mathbf{s} \rangle_n \mathbf{x}_n^T \right], \quad a_2 = \frac{1}{ND} \sum_{n=1}^N \text{tr} \left[ \langle \mathbf{s} \mathbf{s}^T \rangle_n \right]. \quad (4.5)$$

Analytic expressions for roots of a cubic equation exist (see, e.g., Press, Flannery, Teukolsky, & Vetterling, 1988).

For both the second and the third terms of equation 4.1, we may observe that they can be written as a sum over  $D$  independent terms:

$$Q_2(\tilde{\mu}, \tilde{\sigma} | \mu, \sigma) = \frac{1}{N} \sum_{n=1}^N \sum_{i=1}^D \int ds_i \sum_{z_i=1}^{M_i} p(s_i, z_i | \mathbf{x}_n) \log p(s_i | z_i) \quad (4.6)$$

$$Q_3(\tilde{\pi} | \pi) = \frac{1}{N} \sum_{n=1}^N \sum_{i=1}^D \sum_{z_i=1}^{M_i} p(z_i | \mathbf{x}_n) \log p(z_i). \quad (4.7)$$

Using equation 3.10, we can derive the following update rules for  $\tilde{\pi}$ ,  $\tilde{\mu}$ , and  $\tilde{\sigma}$ ,

$$\tilde{\pi}_{z_i} = \frac{1}{N} \sum_{n=1}^N \alpha_{z_i}^n, \quad \tilde{\mu}_{z_i} = \frac{\sum_{n=1}^N \alpha_{z_i}^n \langle s_i \rangle_{z_i}^n}{\sum_{n=1}^N \alpha_{z_i}^n}, \quad \tilde{\sigma}_{z_i}^2 = \frac{\sum_{n=1}^N \alpha_{z_i}^n \langle s_i^2 \rangle_{z_i}^n}{\sum_{n=1}^N \alpha_{z_i}^n} - \tilde{\mu}_{z_i}^2, \quad (4.8)$$

where  $\langle \cdot \rangle_{z_i}^n$  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  $\langle s_i \rangle_{z_i}^n$ ,  $\langle s_i^2 \rangle_{z_i}^n$ ,  $\langle s_i \rangle_n$ , and  $\langle s_i^2 \rangle_n$ . Notice that we do not have to calculate the off-diagonal terms of the covariances between the sources given the data,  $\langle s_i s_j \rangle_n$ . The former statistics can be computed using equations 3.8 through 3.10:

$$\langle s_i \rangle_{z_i}^n = b_{z_i}^n, \quad \langle s_i \rangle_n = \sum_{z_i=1}^{M_i} \alpha_{z_i}^n b_{z_i}^n \quad (4.9)$$

$$\langle s_i^2 \rangle_{z_i}^n = \gamma_{z_i}^2 + (b_{z_i}^n)^2, \quad \langle s_i^2 \rangle_n = \sum_{z_i=1}^{M_i} \alpha_{z_i}^n (\gamma_{z_i}^2 + (b_{z_i}^n)^2). \quad (4.10)$$

**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 as to whether we can use simpler source models. The method advocated in the extended infomax algorithm (Lee, Girolami, & Sejnowski, 1999) is to switch between a sub- and a supergaussian model depending on the sign of a switching moment, which tests the stability of the solution (Cardoso & Laheld, 1996). This method requires the estimation of only 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 supergaussian or a subgaussian mixture. The same behavior 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 supergaussian and a probability  $1 - r$  that it is subgaussian. 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 supergaussian density and a set that is able to model a subgaussian density:

$$\pi_{z_i} = \begin{cases} r_i \rho_{z_i} & z_i \in [1, \dots, K_i] \\ (1 - r_i) \rho_{z_i} & z_i \in [K_i + 1, \dots, M_i]. \end{cases} \quad (4.11)$$

The values of  $\rho_{z_i}$  are fixed and chosen such that for every  $i$ , the mixture coefficients with  $z_i = 1, \dots, K_i$  describe a supergaussian density and the coefficients with  $z_i = K_i + 1, \dots, M_i$  describe a subgaussian 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

$$\tilde{r}_i = \frac{1}{N} \sum_{n=1}^N \sum_{z_i=1}^{K_i} \alpha_{z_i}^n. \quad (4.12)$$

For every source, the algorithm will estimate the optimal mixture of these two model densities. For supergaussian sources, the value of  $r$  is higher than for subgaussian 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  $\beta$  has a significant value, this procedure is suboptimal. A better way is to consider the

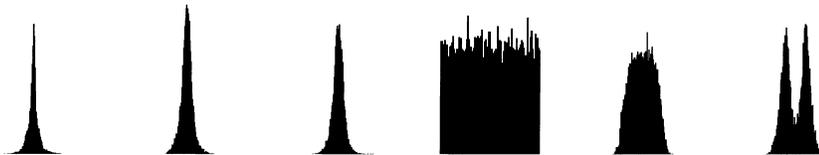


Figure 1: Histograms of data from three sound sources (first three) and three artificial sources (last three).

maximum a posteriori (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,

$$\begin{aligned} \hat{s}_{in} &= \arg \max_{s_i} p(s_i | \mathbf{x}_n) \\ &= \arg \max_{s_i} \sum_{z_i=1}^{M_i} \alpha_{z_i}^n \mathcal{G}_{s_i} [b_{z_i}^n, \gamma_{z_i}^2], \quad i = 1, \dots, D, \end{aligned} \quad (4.13)$$

where we used equation 3.10. 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 three sound sources<sup>2</sup> together with three artificial sources. The sound data had a supergaussian profile, and 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 1 s on the diagonal and 0.25 in the off-diagonal entries. No noise was added to the mixed sources. The algorithm was run in its fully adaptive mode (two gaussians per source) and the soft-switch mode (a mixture of two gaussian mixtures consisting of two gaussians each). As the 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<sup>3</sup> between the true and estimated mixing matrix, the log-likelihood, the noise variance, and the mixing coefficients (soft switch only). The results for the fully adaptive algorithm (dashed lines) and the soft-switch approach (solid lines) are very similar, although

<sup>2</sup> The CD recordings can be found online at <http://sweat.cs.unm.edu/bap/demos.html>.

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

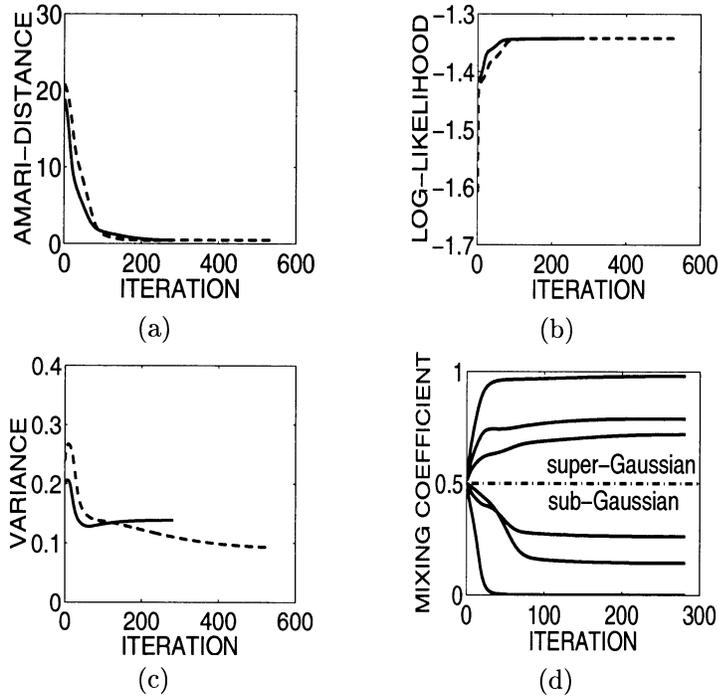


Figure 2: Evolution of (a) Amari distance, (b) log-likelihood, (c) noise variance, and (d) mixing coefficients over time. Solid lines indicate the results for the soft-switch algorithm; dashed lines show the results for the fully adaptive approach.

convergence was faster in the soft-switch mode. Furthermore, the estimated noise variance in the fully adaptive mode was lower due to the greater flexibility in modeling the source densities. From the mixing coefficients (see Figure 2d) it can be seen that three supergaussian (sound) sources and three subgaussian (artificial) sources were recovered. The estimated mixing matrices for both approaches, after correcting for scale and permutation, are given by:

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

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

## 6 Discussion and Conclusion

---

This article proposes using the EM algorithm to perform ICA. The source vector  $\mathbf{s}$  is considered as a hidden variable that 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, and 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 the focus is on source separation (instead of density estimation), a simpler algorithm was developed that uses a mixture of a sub- and a supergaussian model for the source densities. One parameter per source, measuring the probability that the data in that direction are explained by the supergaussian model (as opposed to being explained by the subgaussian 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 important, it allows factoring 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 need to calculate only two simple sufficient statistics: the source mean and the source variance given the data. Calculating 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 estimate only a square mixing matrix and are restricted to modeling 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

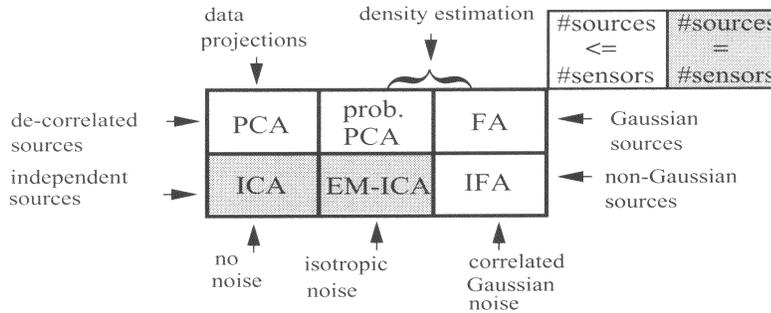


Figure 3: Table of techniques for data analysis. The box contains six methods that relate the observed data linearly to hidden sources. From left to right, top to bottom, they are: principal component analysis, probabilistic principal component analysis, factor analysis, independent component analysis, EM-independent component analysis, and independent factor analysis. Methods in the 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.

problem. First, if the interesting directions contain most of the power of the signal, PCA can be used to 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, say, 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 by 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.

Figure 3 shows the difference between EM-ICA and similar related methods. It lists several data analysis techniques that 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 & Bishop, 1997), a simple isotropic noise model is used to capture the variance in the directions orthogonal to the principal components. Factor analysis is obtained when more general correlated gaussian noise is allowed. The bottom row still relates the sources linearly to the observed data but uses a nongaussian factorized distribution for these sources. While standard ICA searches for independent projections of the data, EM-ICA fits a probability density; that is, alongside the estimation of a mixing matrix, 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 (IFA) (Attias, 1999) is an extension of EM-ICA where more general gaussian noise is modeled. An important property that IFA shares with PCA, probabilistic PCA, and factor analysis, but not with ICA and EM-ICA, is that the number of sources does not have to match the number of sensors. The disadvantage of IFA is 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).

### Acknowledgments

---

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

### References

---

- Amari, S., Cichocki, A., & Yang, H. (1996). A new algorithm for blind signal separation. In D. Touretzky, M. Mozer, and M. Hasselmo (Eds.), *Advances in neural information processing systems*, 8 (pp. 757–763). Cambridge, MA: MIT Press.
- Attias, H. (1999). Independent factor analysis. *Neural Computation*, 11, 803–851.
- Barlow, H. (1989). Unsupervised learning. *Neural Computation*, 1, 295–311.
- Bell, A., & Sejnowski, T. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7, 1129–1159.
- Bell, A., & Sejnowski, T. (1997). The independent components of natural scenes are edge filters. *Vision Research*, 37, 3327–3338.
- Cardoso, J., & 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., & 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* (Tech. Rep.). Helsinki University of Technology, Laboratory of Computer and Information Science.

- Jutten, C., & Héroult, J. (1991). Blind separation of sources, part 1: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 2, 1–10.
- Lee, T., Girolami, M., & 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., & Sejnowski, T. (2000). Learning overcomplete representations. *Neural Computation*, 12, 337–366.
- Makeig, S., Bell, A., Jung, T., & 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., & 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., & Gassiat, E. (1997). Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *Proc. ICASSP*, 5, 3617–3620.
- Oja, E. (1997). The nonlinear PCA learning rule in independent component analysis. *Neurocomputing*, 17, 25–45.
- Pearlmutter, B., & Parra, L. (1996). A context sensitive generalization of ICA. In *International Conference on Neural Information Processing* (pp. 151–157).
- Press, W., Flannery, B., Teukolsky, S., & Vetterling, W. (1988). *Numerical recipes in C*. Cambridge: Cambridge University Press.
- Tipping, M., & Bishop, C. (1997). *Probabilistic principal component analysis* (Tech. Rep. No. NCRG/97/010). Aston University, Department of Computer Science and Applied Mathematics, Neural Computing Research Group.
- Welling, M., & Weber, M. (1999). Independent component analysis of incomplete data. In *Proceedings of the 6th Joint Symposium on Neural Computation*, 9, 162–168.

---

Received March 15, 1999; accepted June 1, 2000.