FAST ALGORITHMS FOR BAYESIAN INDEPENDENT COMPONENT ANALYSIS

Harri Valpola and Petteri Pajunen

Helsinki University of Technology
Lab. of Computer and Information Science
P.O. Box 5400
FIN-02015 HUT
Finland

ABSTRACT

Fast algorithms for linear blind source separation are developed. The fast convergence is first derived from low-noise approximation of the EM-algorithm given in [1], to which a modification is made that leads as a special case to the FastICA algorithm [2]. The modification is given a general interpretation and is applied to Bayesian blind source separation of noisy signals.

1. INTRODUCTION

We consider the problem of finding linearly mixed source signals $\mathbf{s}(t) = [s_1(t), \ldots, s_m(t)]^T$ from observed noisy linear mixtures

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

The mixing matrix $\mathbf{A}$ is unknown and $\mathbf{n}(t)$ is additive noise. When we consider a finite number of observed mixture samples, we may write the data model in matrix form as

$$\mathbf{X} = \mathbf{A}\mathbf{S} + \mathbf{N}$$

Each time-indexed matrix contains a sequence of vector samples, e.g.

$$\mathbf{S} = [\mathbf{s}(1), \mathbf{s}(2), \ldots, \mathbf{s}(M)]$$

To find the mixing matrix $\mathbf{A}$, it is necessary that the source signals possess certain statistical properties. For example, it is sufficient that the source signals are not Gaussian. It is also sufficient that the possibly Gaussian sources have time dependencies, together with some other conditions.

2. EM-ALGORITHM FOR INDEPENDENT COMPONENT ANALYSIS

In the signal model only the vectors $\mathbf{x}(t)$ are observed. Everything else is unknown and must be estimated using the data. In general, the task is to compute the joint posterior distribution for all the unknown parameters conditioned by the mixtures $\mathbf{x}(t)$.

A more simple case is when the maximum likelihood estimate is used for some on the parameters. This can be done by the EM-algorithm where the computation alternates between computing the posterior distribution of one set of variables given the current point estimate of the other set of variables (E-step) and then using the posterior distribution of the first set of variables to compute a new maximum likelihood estimate of the second set of variables (M-step).

When EM-algorithm is applied to ICA, usually the full posterior distribution is computed for sources and the maximum likelihood estimate is used for the rest of the parameters. This means that in the E-step we need to compute the posterior distribution of the sources $\mathbf{s}$ given $\mathbf{x}$,$\mathbf{A}$ and the noise covariance $\sigma^2\mathbf{I}$

$$p(\mathbf{s}|\mathbf{A}, \mathbf{x}, \sigma^2\mathbf{I})$$

and use it to update our estimates.

Using the matrix notation for the finite number of samples, i.e. $\mathbf{X}$ and $\mathbf{S}$, we can write the M-step (see [3]) re-estimation for the mixing matrix as

$$\hat{\mathbf{A}} = \mathbf{R}_{xs}\mathbf{R}_{ss}^{-1}$$

where the posterior correlation matrices are

$$\mathbf{R}_{xs} = \frac{1}{M}\sum_i \mathbf{x}(i)\,\mathrm{E}\{\mathbf{s}^T(i)|\mathbf{x}(i), \mathbf{A}, \sigma^2\mathbf{I}\} = \mathbf{X}\hat{\mathbf{S}}^T/M$$

$$\mathbf{R}_{ss} = \frac{1}{M}\sum \mathrm{E}\{\mathbf{s}(i)\mathbf{s}^T(i)|\mathbf{x}(i), \mathbf{A}, \sigma^2\mathbf{I}\} = \widehat{\mathbf{S}\mathbf{S}}^T/M\,.$$

The expectations are taken over the posterior distribution of the sources.

We will consider here the case where σ^2 is small. If we further assume that the mixtures are prewhitened,

we can constrain the mixing matrix to be orthogonal and we can assume that the sources have unit variance. This makes $\mathbf{R}_{ss}$ a unit matrix.

In [1] the EM-algorithm is derived as a low-noise approximation for the case of square mixing matrix $\mathbf{A}$. First, the posterior mean $p(\mathbf{s}|\mathbf{A},\mathbf{x},\sigma^2\mathbf{I})$ is obtained as

$$\hat{\mathbf{s}} = \mathrm{E}\{\mathbf{s}|\mathbf{A},\mathbf{x},\sigma^2\mathbf{I}\} \approx \mathbf{s}_0 + \sigma^2(\mathbf{A}^T\mathbf{A})^{-1}f(\mathbf{s}_0)$$

where $f(\cdot)$ is the derivative $\frac{\partial \log p_i(s_i)}{\partial s_i}$ and $\mathbf{s}_0 = \mathbf{A}^{-1}\mathbf{x}$. Since we assumed that the mixing matrix is orthogonal, we can omit the term $(\mathbf{A}^T\mathbf{A})^{-1}$ and we get

$$\hat{\mathbf{s}} = \mathrm{E}\{\mathbf{s}|\mathbf{A},\mathbf{x},\sigma^2\mathbf{I}\} \approx \mathbf{s}_0 + \sigma^2 f(\mathbf{s}_0)$$

Substituting the above approximations we get

$$\begin{aligned}\hat{\mathbf{A}} &= \mathbf{X}\hat{\mathbf{S}}^T/M \\ &\approx \mathbf{X}\mathbf{S}_0^T/M + \sigma^2\mathbf{X}\mathbf{F}(\mathbf{S}_0^T)/M \\ &= \mathbf{A} + \sigma^2\mathbf{X}\mathbf{F}(\mathbf{S}_0^T)/M\end{aligned}$$

As the authors mention in [1], this approximation leads to an EM-algorithm which converges slowly with low noise variance σ^2. They also point out that there is no visible "noise-correction". It is precisely this point that we will address in the next section.

3. FAST EM-ALGORITHM BY FILTERING OF GAUSSIAN NOISE

With low noise variance σ^2 the convergence of the EM-algorithm to the optimal value takes a time proportional to $1/\sigma^2$. We will next show how the re-estimation step can be modified so that the convergence rate will be independent of σ^2 which yields a significant speedup if σ^2 is small.

Consider the case that we estimate the sources one at a time and that the sources are assumed to be whitened and the mixing matrix $\mathbf{A}$ orthonormal. Denote one of the source signals in the optimal solution as $\hat{\mathbf{s}}_{opt}$. By optimality we mean that the standard EM-algorithm will eventually converge to $\hat{\mathbf{a}}_{opt} = \mathbf{X}\hat{\mathbf{s}}_{opt}^T$ with $\hat{\mathbf{s}}_{opt} = \mathrm{E}\{\mathbf{s}|\hat{\mathbf{a}}_{opt},\mathbf{X},\sigma^2\mathbf{I}\}$.

When we have not yet found the optimal vector $\hat{\mathbf{a}}_{opt}$, we have

$$\mathbf{s}_0 = \alpha\mathbf{s}_{opt} + \beta\mathbf{s}_G$$

where $\alpha^2 + \beta^2 = 1$. The noise $\mathbf{s}_G$ is mostly due to the other sources and to a small extent the Gaussian noise in the data. We can think that the E-step filters away the noise by making use of the knowledge about the prior distribution of $\mathbf{s}$. This gives one point of view into the slow convergence: in low noise case most of the unwanted signal $\mathbf{s}_G$ is due to other sources and a slow convergence results. From this point of view, it is obvious that we can speed up the convergence if we can filter away also that part of $\mathbf{s}_G$ which is due to other sources.

When we are far from the optimal solution, it is natural to assume that $\beta \approx 1$ and $\alpha \approx 0$. Since $\mathbf{a}$ and $\mathbf{s}_0$ are linearly related, we get

$$\alpha\hat{\mathbf{a}}_{opt} = \hat{\mathbf{a}} - \beta\hat{\mathbf{a}}_G \approx \hat{\mathbf{a}} - \hat{\mathbf{a}}_G.$$

If we can compute $\mathbf{a}_G$, we can adjust the vector $\mathbf{a}$ to take into account the apparent noise due to other sources. By the central limit theorem, the distribution of the sum of contributions from several other sources approaches Gaussian as the number of other sources increases. This leads to the following modification: we may estimate $\hat{\mathbf{a}}_G$ using the same re-estimation whose result will be approximately

$$\hat{\mathbf{a}}_G \approx \mathbf{a} + \sigma^2\mathbf{X}_G\mathbf{F}(\mathbf{s}_{0G})/M.$$

where $\mathbf{X}_G$ is the set of mixtures replaced by Gaussian noise with the same covariance as $\mathbf{X}$ and $\mathbf{s}_{0G}$ is the source obtained as $\mathbf{a}^T\mathbf{X}_G$. The Gaussian source $\mathbf{s}_{0G}$ is the projection of Gaussian noise to the subspace spanned by $\mathbf{a}$ and therefore represents the contribution of the other sources and some Gaussian noise to the estimated source $\mathbf{s}_0$. As derived above, we can eliminate much of this noise by updating $\mathbf{a}$ using the difference $\hat{\mathbf{a}} - \hat{\mathbf{a}}_G$ which is then normalized. The normalization can be done, since scaling of the sources is an undeterminacy in ICA.

Taking the difference yields approximately

$$\hat{\mathbf{a}} - \hat{\mathbf{a}}_G \approx \sigma^2[\mathbf{X}\mathbf{F}(\mathbf{s}_0)^T - \mathbf{X}_G\mathbf{F}(\mathbf{s}_{0G})^T]/M$$

which shows that the normalization cancels the effect of σ^2 from the learning rule:

$$\hat{\mathbf{a}}_{new} = \frac{\hat{\mathbf{a}} - \hat{\mathbf{a}}_G}{\|\hat{\mathbf{a}} - \hat{\mathbf{a}}_G\|}.$$

We assumed above that there was a lot of Gaussian noise by approximating $\beta \approx 1$. It turns out that the above modification does not affect the optimal solutions of the algorithm, i.e., if $\hat{\mathbf{a}}_{opt}$ is a fixed point of the original EM-algorithm, it is also a fixed point of the modified algorithm. This follows immediately from the fact that $\hat{\mathbf{a}}_G$ is always parallel to $\mathbf{a}$ since $\mathbf{X}_G$ is spherically symmetric. To get a rough idea about why this is so, suppose there is a vector $\mathbf{b}$ which is orthogonal to $\mathbf{a}$, i.e., $\mathbf{b}^T\mathbf{a} = 0$. Then

$$\mathbf{b}^T\hat{\mathbf{a}}_G \approx \mathbf{b}^T\mathbf{a} + \mathbf{b}^T\mathbf{X}_G\mathbf{F}(\mathbf{s}_{0G}) = \mathbf{s}'_{0G}\mathbf{F}(\mathbf{s}_{0G}) = 0\,.$$

The last step follows from the fact that $\mathbf{s}'_{0G}$ is a projection to an orthogonal direction form $\mathbf{s}_{0G}$ and by Gaussianity of $\mathbf{X}_G$, statistically independent form $\mathbf{F}(s_{0G})$. But since this must hold for all $\mathbf{b}$ which are orthogonal to $\mathbf{a}$, it follows that $\hat{\mathbf{a}}_G$ has to be parallel to $\mathbf{a}$.

In the next section we add validity to the result by showing that FastICA algorithm follows from this procedure.

4. FASTICA AS EM-ALGORITHM WITH FILTERING OF GAUSSIAN NOISE

The FastICA algorithm [2] can be interpreted as performing the above described noise removal. In FastICA the requirement of whitening the sources is also made and therefore $\mathbf{R}_{ss} = \mathbf{I}$ and $(\mathbf{A}^T\mathbf{A})^{-1} = \mathbf{I}$. Then, the sources can be found one by one and we can consider a single column $\mathbf{a}$ of the mixing matrix $\mathbf{A}$.

To derive the FastICA algorithm from the modified EM-algorithm, it is sufficient to note that the term $\mathbf{X}_G\mathbf{F}(\mathbf{s}_{0G})^T/M = \mathbf{a}\mathbf{s}_{0G}F(\mathbf{s}_{0G})^T/M$ is $C_f\mathbf{a}$ where C_f is a constant that depends only on the nonlinear function $f(\cdot)$. Then the update rule is

$$\hat{\mathbf{a}} - \hat{\mathbf{a}}_G = \mathbf{X}\mathbf{F}(\mathbf{s}_0^T) - C_f\mathbf{a}$$

$$\hat{\mathbf{a}}_{new} = \frac{\hat{\mathbf{a}} - \hat{\mathbf{a}}_G}{\|\hat{\mathbf{a}} - \hat{\mathbf{a}}_G\|}$$

which is the FastICA algorithm, where the constant C_f is the expectation $\mathrm{E}\{s_{0G}f(s_{0G})\}$.

The choice of fixed nonlinearity $f(\cdot)$ is implicitly connected to the distribution of the sources s. The derivation of the EM-algorithm required that

$$f(s) = \frac{\partial \log p(s)}{\partial s}$$

However, we see that $f(\cdot)$ has certain degrees of freedom due to taking the difference $\mathbf{X}\mathbf{F}(\mathbf{s}_0^T) - \mathbf{X}_G\mathbf{F}(\mathbf{s}_{0G}^T)$. Expanding f polynomially we obtain $p(s) = \exp(a + bs + cs^2 + dg(s))$ where $g'(s) = f(s)$ and $g(s)$ contains all the powers of f higher than two and possibly lower moments too. This representation follows since in the update rule constants and linear terms of $f(\cdot)$ will cancel out. Therefore they will appear in the distribution $p(s)$ in the exponent with the power raised by one due to integration. Since $p(s)$ must be a probability density, the constant a will be fixed by the requirement $\int p(s)ds = 1$. Mean and variance of s will determine the constants b and c, since the sources are required to be zero-mean and whitened (variance is fixed to unity). There is one free parameter d left, which means that there is not only one distribution corresponding to $f(\cdot)$ but a family of $p(s)$. Typically the family includes both super- and sub-Gaussian densities, which is why the same $f(\cdot)$ can be used for both cases.

5. APPLICATION TO GENERAL ICA ALGORITHMS

The procedure giving faster convergence derived in previous sections is a general approach and FastICA was seen to be a special case. Since the faster convergence was achieved by comparing the re-estimation step to Gaussian noise removal, the approach is valid for any situation where the general noisy ICA model holds with Gaussian noise and linear mixtures. It is not required that the E-step uses the approximation $\hat{\mathbf{s}} \approx \mathbf{s}_0 + \sigma^2 f(\mathbf{s}_0)$; instead, it can be any method that can use $\mathbf{s}_0$ to compute $\hat{\mathbf{s}}$. Denote this estimation by

$$\hat{\mathbf{s}} = \mathbf{g}(\mathbf{s}_0)\,.$$

Then it is always possible to replace the source with Gaussianized source $\mathbf{s}_{0G}$ and obtain

$$\hat{\mathbf{s}}_G = \mathbf{g}(\mathbf{s}_{0G})\,.$$

Having estimated two sets of sources, we can apply any method whatsoever to estimate the mixing matrix using the newly estimated sources. This gives us two new estimates of the vectors $\hat{\mathbf{a}}$ and $\hat{\mathbf{a}}_G$ of the mixing matrix. The final estimate is obtained as the normalized difference as above.

6. APPLICATION TO BAYESIAN NOISY ICA

Above, noise was assumed to have a small variance to justify certain approximations. Therefore the result was not strictly an algorithm for noisy ICA since the approximations get worse with increasing noise variance. Below, we will consider a speedup modification for Bayesian noisy ICA. The Bayesian approach adopted here gives certain important advantages:

- noise can have a finite variance
- source densities need not be fixed a priori; they can be estimated
- the number of sources can be estimated
- model comparison is possible

Specifically the source distributions are modeled as mixtures of Gaussians and the posterior is approximated using ensemble learning. The treatment of the source distribution is similar to [4] which uses a factorial approximation of the posterior source distributions in connection with the EM-algorithm. The modification would be directly applicable to the algorithm in [4], but we will consider an algorithm where all the posterior distribution is estimated for all parameters, i.e., point estimates are not used at all.

6.1. Overview of the Bayesian ICA Algorithm

In [5], it is described how to use ensemble learning for the noisy ICA model. The posterior distribution is over all unknown parameters, including the mixing matrix $\mathbf{A}$. In ensemble learning, a factorial approximation $q(\mathbf{S}, \mathbf{A}, \dots)$ is fitted to the actual posterior distribution $p(\mathbf{S}, \mathbf{A}, \dots | \mathbf{X})$ by minimising the Kullback-Leibler information between them, i.e., the cost function which is minimized during learning is

$$I(q; p) = \mathrm{E}_q\{\log(q/p)\}\,.$$

The algorithm is computationally efficient when the approximation $q(\cdot)$ of the posterior probability $p(\cdot|\mathbf{X})$ is chosen to be factorial. This can be seen as an extension of the factorial EM-algorithm in [4], where $q(\cdot)$ included only the posterior distribution of the sources. For further details, see for instance [6, 7], where ensemble learning is applied to nonlinear ICA.

The ICA algorithm based on ensemble learning works in much the same way as EM-algorithm. First the distribution of the sources is computed by using the current estimate of the distribution of the other parameters. Then the distribution of the other parameters is computed using this distribution of the sources. The posterior distributions of the parameters are approximated by Gaussian distribution which means that for each element of the mixing matrix $\mathbf{A}$, the posterior mean and variance is estimated. The modification will be applied to the posterior mean of the mixing matrix.

For each vector of the mixing matrix, the modified posterior mean will be the normalized difference between the posterior mean estimated from the original sources and the Gaussianized sources. The iteration is then repeated by estimating the posteriors of the sources again, using the new parameter distribution.

In practice, the algorithm is performed in deflatory manner, that is, the sources are extracted one by one. The mixtures are prewhitened and then the mixing matrix is estimated one column $\mathbf{a}$ at a time.

A heuristic stabilization is added to ensure convergence. This is achieved by updating the vector $\mathbf{a}$ to be a linear combination $\alpha\mathbf{a}_{new} + (1-\alpha)\mathbf{a}_{old}$. The coefficient α is increased when consecutive corrections to $\mathbf{a}$ have a positive inner product which means that they do not change to opposite directions. Otherwise, α is decreased.

7. EXPERIMENTS

The Bayesian ICA algorithm was tested on MEG data, which is identical to the data used in [8]. The data has 122 channels of measurements over two minutes digitized at 148 Hz. The measurements contain signals resulting in the electrical activity of the brain but also signals which can be considered artifacts. These include signals caused by muscular activity, eye movements, cardiac rhythm and even a signal caused by a digital watch that the test subject was wearing. Since it can be assumed that most of the artifacts are independent of the brain activity, it is hoped that ICA can find the artifacts.

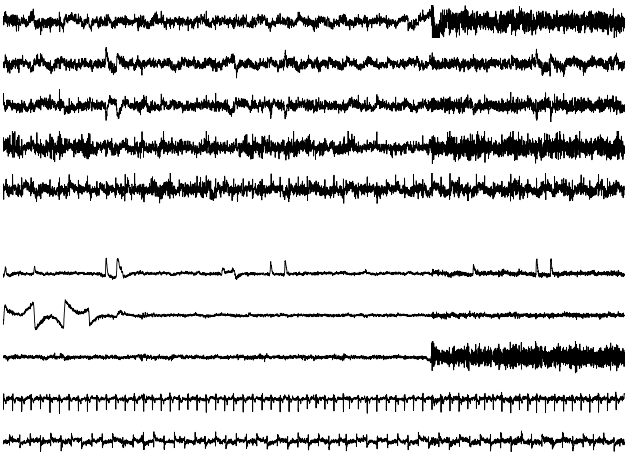

Figure 1: Above: five MEG measurements. Below: five separated sources found by the Bayesian ICA algorithm.

The Bayesian ICA algorithm was used to separate 30 sources from the 122 measures channels. The results obtained were comparable to those reported in [8]. In figure 1, five measurements and five non-Gaussian sources found by the algorithm are illustrated.

The modification of the estimate of $\mathbf{a}$ by the estimate $\mathbf{a}_G$ typically reduces the convergence time by a factor of ten; the iteration typically converged in 30 iterations.

8. DISCUSSION

First we considered the EM-algorithm for finding independent components with low noise. The problem of slow convergence was noted and an improvement was proposed. When finding the sources one at a time, the contributions of the unwanted sources was treated as noise, which leads to faster convergence. Although the approach was found to be implicitly the same as in the FastICA algorithm, it is valid for other situations too. In Bayesian ICA for i.i.d. sources, the modification can be applied as proposed. Other possibilities include finding groups of components that are not mutually independent but are independent related to other compo-

nents not in the group. The independent components are then projections to multidimensional subspaces instead of one-dimensional projections. This has been proposed e.g. in [9, 10]. The modification proposed in this paper applies to this case too, since the contributions of sources not in the group can be regarded as approximately Gaussian noise.

Further work includes finding more general principles, where the modification could be derived for other cases, such as time-dependent sources or nonlinear ICA.

9. REFERENCES

[1] O. Bermond and J. Cardoso, "Approximate likelihood for noisy mixtures," in *Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA'99)*, (Aussois, France), pp. 325–330, January 1999.

[2] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Computation*, vol. 9, no. 7, pp. 1483–1492, 1997.

[3] E. Moulines, J. Cardoso, and E. Gassiat, "Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models," in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 3617–3620, 1997.

[4] H. Attias, "Independent factor analysis," *Neural Computation*, vol. 11, no. 4, pp. 803–851, 1999.

[5] H. Lappalainen, "Ensemble learning for independent component analysis," in *Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA'99)*, (Aussois, France), pp. 7–12, January 1999.

[6] H. Valpola, "Nonlinear independent component analysis using ensemble learning: Theory," in *Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA2000)*, (Helsinki, Finland), June 2000.

[7] H. Valpola, X. Giannakopoulos, A. Honkela, and J. Karhunen, "Nonlinear independent component analysis using ensemble learning: Experiments and discussion," in *Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA2000)*, (Helsinki, Finland), June 2000.

[8] R. Vigario, V. Jousmäki, M. Hämäläinen, R. Hari, and E. Oja, "Independent component analysis for identification of artifacts in magnetoencephalographic recordings," in *Advances in Neural Information Processing Systems 10 (NIPS'97)*, (Cambridge, MA), pp. 229–235, MIT Press.

[9] J.-F. Cardoso, "Multidimensional independent component analysis," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'98)*, vol. 4, (Seattle, Washington, USA), pp. 1941–1944, May 1998.

[10] A. Hyvärinen and P. O. Hoyer, "Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces," *Neural Computation*, 2000. (in press).