Flexible and efficient implementations of Bayesian independent component analysis

Winther, Ole; Petersen, Kaare Brandt

Published in:
Neurocomputing

Link to article, DOI:
10.1016/j.neucom.2007.01.007

Publication date:
2007

Document Version
Early version, also known as pre-print

Elsevier Editorial System(tm) for Neurocomputing

Manuscript Draft

Manuscript Number: NEUCOM-D-05-00490R1

Title: Flexible and Efficient Implementations of Bayesian Independent Component Analysis

Article Type: Full Length Article (FLA)

Section/Category:

Keywords: Independent Component Analysis, Empirical Bayes, Mean Field Methods, Variational methods

Corresponding Author: Mr Kaare Brandt Petersen, PhD

Corresponding Author's Institution: Technical University of Denmark

First Author: Ole Winther, PhD

Order of Authors: Ole Winther, PhD; Kaare Brandt Petersen, PhD

Manuscript Region of Origin:

Abstract: In this paper we present an empirical Bayes method for flexible and efficient Independent Component Analysis (ICA). The method is flexible with respect to choice of source prior, dimensionality and positivity of the mixing matrix, and structure of the noise covariance matrix. The efficiency is ensured using parameter optimizers which are more advanced than the expectation maximization (EM) algorithm, but still easy to implement. These optimizers are the overrelaxed adaptive EM algorithm and the easy gradient recipe. The required expectations over the source posterior are estimated with accurate mean field

methods: variational and the expectation consistent framework.

We demonstrate the usefulness of the approach with the publicly available Matlab toolbox icaMF.

Flexible and Efficient Implementations of Bayesian Independent Component Analysis

Ole Winter, Kaare Brandt Petersen

Technical University of Denmark, Building 321, DK2800 Kongens Lyngby, Denmark Europe

Abstract

In this paper we present an empirical Bayes method for flexible and efficient Independent Component Analysis (ICA). The method is flexible with respect to choice of source prior, dimensionality and constraints of the mixing matrix (positivity), and structure of the noise covariance matrix. The efficiency is ensured using parameter optimizers which are more advanced than the expectation maximization (EM) algorithm, but still easy to implement. These optimizers are the overrelaxed adaptive EM algorithm and the easy gradient recipe. The required expectations over the source posterior are estimated with mean field methods: variational and the expectation consistent (EC) framework. We describe the derivation of the EC framework for ICA in detail and give empirical results demonstrating the improved performance. We demonstrate the usefulness of the approach with the publicly available Matlab toolbox `icaMF`.

Key words: Independent Component Analysis, Empirical Bayes, Mean Field Methods, Variational methods
PACS:

1 Introduction

Since Independent Component Analysis (ICA) in the early nineties caught the attention of the machine learning community, the interest and activities within this area have all but exploded. Although initially regarded as an example of a blind source separation problem for independent data, the focus has in recent years gradually shifted from different aspects of this instantaneous problem to the challenge of the convolutive case (mixing over time). A process fuelled by algorithms such as InfoMax [2] and FastICA [7] which, although not very flexible, are robust and fast.

But the completely general instantaneous case is far from solved: So far there exists no algorithm which can do noisy, non-square mixing with an arbitrary non-gaussian prior with the same robustness and speed as the above mentioned algorithms. Variational (so-called mean field and also known as ensemble learning) methods [8,1,23,11,4,6] are attractive because they are very flexible general modelling tools. The mean field ICA method as described in [6], however, had two major difficulties: First, the flexibility with respect to the prior makes the inside of the black-box rather complicated and unattractive to a wide range of the application-driven part of the research community. Second, it was slow to converge and no universal stopping criteria could be given. These two difficulties, however, can now be handled, as this paper demonstrates: The complexity by the availability of a Matlab toolbox with plug-and-play demos and examples and the convergence by efficient optimization schemes beyond the traditional expectation maximization (EM) algorithm .

In a broader context reaching well beyond ICA, the mean field methods such as variational Bayes, loopy belief propagation, expectation propagation (EP) and expectation consistent (EC) have recently gathered much interest, see e.g. [8,1,15,25,16,10,17] because of their potential as approximate Bayesian inference engines. An undesirable property of the mean field methods in this context is that the approximation error is unattainable. One cannot in any quantitative manner say much about the deviation of marginal moments or likelihoods from their true values. But in many tests, however, the accuracy and the polynomial complexity of the mean field methods to intractable sums and integrals, is positioning the mean field methods as a high-end approximation.

In this paper we give for the first time an application of the EC framework and EP message passing algorithm to the ICA model. EC and EP are closely related to the adaptive TAP framework [15,16,6,5]. In fact these non-linear iterative methods share fixed points. Whereas the adaptive TAP fixed point equations were derived from the so-called cavity method—a central limit theorem argument used to derive an tractable approximations to the marginal and predictive distributions—EC is a formalization of the same underlying idea aiming at giving an approximation to the marginal likelihood. A set of complementary variational distributions, which share sufficient statistics, arise naturally in the framework. EP is an intuitively appealing scheme for tuning the variational distributions to achieve this expectation consistency. We will not go into detail with cavity derivation here since, although it sheds lights on why the approximation will work, nevertheless is less formalized than EC. The interested reader is referred to Refs. [16,5] and references therein.

For the last few years, it has been observed that an important cause for the non-robustness of the mean field methods rests not only upon the approximation error but rather on the slow convergence of the expectation maximization

(EM) style algorithms, typically used for the joint parameter-latent variable inference problem [20]. Several alternatives to EM learning, also applicable to non-mean field based inference, have been proposed recently [22,14] and analysis have been given to explain convergence failure both in general and specific settings [21]. Furthermore, these methods in some cases make the difference between convergence in finite time or not [18], but in most cases they at least give a huge speed-up. This new insight has thus taken the mean field methods one step closer to realizing their full potential for Bayesian inference.

In this paper we revisit mean field ICA to demonstrate that the newly found insights can give us a much more efficient system while still retaining the flexibility of the Bayesian ICA approach. Compared to methods like InfoMax and Fast ICA, the flexibility is rather extensive and enables the user to handle both over-/underdetermined and square mixing, positive constraints on the mixing matrix, noise estimation and general source priors, e.g. positive or discrete. The contributions are three-fold: Application of the EC framework and EP message passing in the ICA context, using precise approximate inference and EM type optimization to give an efficient and flexible framework for ICA and a freely available software package implementation. The paper is organized as follows: Section 2 formulates the instantaneous noisy ICA problem which is the basic model of interest. Section 3 explains how the given log likelihood (bound) is optimized giving three different approaches. Section 4 deals with estimation of the source statistics. Different mean field methods – expectation consistent and variational – are applied to the ICA model and in Section 5 we present the Matlab toolbox. Finally in Section 6 and 7 we wrap up with demonstrations of the developed methods and conclusions.

2 Instantaneous ICA

In this section, we give a quick recap of the empirical Bayes approach to instantaneous ICA with additive Gaussian noise – for a more detailed account the reader is referred to e.g. [6]. The observation model is given by

$$\mathbf{x}_t = \mathbf{A}\mathbf{s}_t + \mathbf{n}_t, \qquad t = 1, ..., N$$

with N being the number of samples. The noise is assumed zero-mean gaussian with covariance $\mathbf{\Sigma}$, i.e. the Likelihood is $p(\mathbf{x}_t|\mathbf{A}, \mathbf{s}_t, \mathbf{\Sigma}) = N(\mathbf{x}_t; \mathbf{A}\mathbf{s}_t, \mathbf{\Sigma})$. The source prior factorizes in both sources and time steps. Denoting the stacked sources by the matrix $\mathbf{S}$, we can write the prior as $p(\mathbf{S}|\boldsymbol{\nu}) = \prod_{it} p_i(S_{it}|\boldsymbol{\nu}_i)$, where $\boldsymbol{\nu}$ is shorthand for the parameters of the prior. The observation vectors $\mathbf{x}_t$ are stacked as columns into one matrix $\mathbf{X}$: $p(\mathbf{X}|\mathbf{A}, \mathbf{S}, \mathbf{\Sigma}) = \prod_t p(\mathbf{x}_t|\mathbf{A}, \mathbf{s}_t, \mathbf{\Sigma})$ and the posterior is given by $p(\mathbf{S}|\mathbf{X}, \boldsymbol{\theta}) = \frac{p(\mathbf{X}|\mathbf{A},\mathbf{S},\mathbf{\Sigma})p(\mathbf{S}|\boldsymbol{\nu})}{p(\mathbf{X}|\boldsymbol{\theta})}$, where we have used the shorthand $\boldsymbol{\theta} = \{\mathbf{A}, \mathbf{\Sigma}, \boldsymbol{\nu}\}$ for the parameters. In the empirical Bayes (or

Maximum Likelihood II) approach applied to ICA, the noise realization and the unobserved sources are integrated out, leaving the parameters $\boldsymbol{\theta}$ to be determined by maximizing the (marginal) Likelihood: $p(\mathbf{X}|\boldsymbol{\theta})$. Alternatively, one may use a hierarchical Bayesian approach [1] marginalizing also over $\boldsymbol{\theta}$, see [11] for an application to ICA.

The log likelihood $\mathcal{L}(\boldsymbol{\theta}) = \ln p(\mathbf{X}|\boldsymbol{\theta})$ is for most priors too complicated for practical approaches and instead a lower bound is used as objective function. The lower bound B is defined by

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) \equiv \ln p(\mathbf{X}|\boldsymbol{\theta}) &= \ln \int q(\mathbf{S}|\boldsymbol{\phi}) \frac{p(\mathbf{X},\mathbf{S}|\boldsymbol{\theta})}{q(\mathbf{S}|\boldsymbol{\phi})} d\mathbf{S} \\ &\geq \int q(\mathbf{S}|\boldsymbol{\phi}) \ln \frac{p(\mathbf{X},\mathbf{S}|\boldsymbol{\theta})}{q(\mathbf{S}|\boldsymbol{\phi})} d\mathbf{S} \equiv B(\boldsymbol{\theta},\boldsymbol{\phi}) \ . \end{aligned} \tag{1}$$

The bounding property is a simple consequence of Jensen's inequality and holds for *any* choice of variational distribution $q(\mathbf{S}|\boldsymbol{\phi})$. In fact it is easy to show that $\mathcal{L}(\boldsymbol{\theta}) = B(\boldsymbol{\theta},\boldsymbol{\phi}) - KL(q,p)$, where $KL(q,p) \geq 0$ denotes the Kullback-Leibler divergence between the variational distribution and the source posterior. Thus, if the variational distribution becomes equal to the source posterior, $KL(p,p) = 0$ and the bound is equal to the log likelihood.

The derivatives of the bound are easily derived for the ICA model

$$\frac{\partial B(\boldsymbol{\theta},\boldsymbol{\phi})}{\partial \mathbf{A}} = \boldsymbol{\Sigma}^{-1} \left(\mathbf{X} \langle \mathbf{S} \rangle_q^T - \mathbf{A} \langle \mathbf{S}\mathbf{S}^T \rangle_q \right) \tag{2}$$

$$\frac{\partial B(\boldsymbol{\theta},\boldsymbol{\phi})}{\partial \boldsymbol{\Sigma}} = \frac{N}{2}\boldsymbol{\Sigma}^{-1} - \frac{1}{2}\boldsymbol{\Sigma}^{-1} \langle (\mathbf{X}-\mathbf{A}\mathbf{S})(\mathbf{X}-\mathbf{A}\mathbf{S})^T \rangle_q \boldsymbol{\Sigma}^{-1} \tag{3}$$

$$\frac{\partial B(\boldsymbol{\theta},\boldsymbol{\phi})}{\partial \boldsymbol{\nu}} = \langle \frac{\partial \ln p(\mathbf{s}|\boldsymbol{\nu})}{\partial \boldsymbol{\nu}} \rangle_q \ . \tag{4}$$

Constrained variables are handled by reparametrization and considered in detail in Section 3.4. Since many different source priors are relevant depending on the application, the derivative involving the prior parameter is not specified in details but left open for easier modular fit to various problem-specific priors.

3 Optimization of Parameters

In this section we discuss a number of approaches for optimization of the lower bound of the log likelihood. The starting point is the EM algorithm for its simplicity but also variants thereof are presented for faster convergence. Note that the convergence issue is not merely a question of convenience but rather a critical part of the overall performance of the approach.

3.1 The EM Algorithm

The traditional optimization applied in [6] is the Expectation-Maximization (EM) algorithm as presented in [12]. In their formulation, the EM algorithm is a coordinate-wise descend in the so-called free energy, which in this context is minus the lower bound function. In short, the EM algorithm for $B(\boldsymbol{\theta}, \boldsymbol{\phi})$ is

E: Maximize $B(\boldsymbol{\theta}, \boldsymbol{\phi})$ with respect to $\boldsymbol{\phi}$ keeping $\boldsymbol{\theta}$ fixed.
M: Maximize $B(\boldsymbol{\theta}, \boldsymbol{\phi})$ with respect to $\boldsymbol{\theta}$ keeping $\boldsymbol{\phi}$ fixed.

This results holds for any variational distribution. However, if the choice of q is constrained to a family of distributions such that the E-step does *not* give $q(\mathbf{S}|\boldsymbol{\phi}) = p(\mathbf{S}|\mathbf{X}, \boldsymbol{\theta})$ then we are only optimizing the bound and not the likelihood itself. This variational approximation works well in many cases, see Section 6.

In the M-step the lower bound $B(\boldsymbol{\theta}, \boldsymbol{\phi})$ is maximized with respect to the model (hyper) parameters $\boldsymbol{\theta} = \{\mathbf{A}, \boldsymbol{\Sigma}, \boldsymbol{\nu}\}$. Setting the derivatives in eqs. (2) and (3) equal to zero, one obtains the following EM updates for the mixing matrix and the noise covariance

$$\mathbf{A} = \mathbf{X}\langle \mathbf{S} \rangle_q^T \langle \mathbf{S}\mathbf{S}^T \rangle_q^{-1} \tag{5}$$

$$\boldsymbol{\Sigma} = \tfrac{1}{N} \langle (\mathbf{X} - \mathbf{A}\mathbf{S})(\mathbf{X} - \mathbf{A}\mathbf{S})^T \rangle_q \ . \tag{6}$$

The corresponding result for the prior parameters cannot be expressed explicitly without choosing what prior to use, and is therefore left for the user of whatever specific prior. With these estimates we are ready to make another E-step with the new values for the parameters, then another M-step, and so on.

The EM algorithm is simple and have important theoretical convergence properties. But it is also sometimes unreasonably slow – a postulate reported and documented by a number of papers regarding the use of EM algorithm for ICA. The results of [20] is based on a small scale statistical investigation, and [3] and [19] provide analytically insight to the experience of slow convergence in the low noise limit.

The analysis in the two latter papers is based on a Taylor expansion of the true source posterior moments in the noise variance. That is, when for simplicity the noise is assumed isotropic with variance σ^2, the source posterior moments $\langle \mathbf{s}_t \rangle$ and $\langle \mathbf{s}_t \mathbf{s}_t^T \rangle$ are expanded in σ^2 and inserted to give the update of the the mixing matrix. The result is

$$\mathbf{A}_{n+1} = \mathbf{A}_n + \mathcal{O}(\sigma^2) \ ,$$

where $\mathbf{A}_n$ denotes the nth estimate of the mixing matrix. This shows that if the noise variance is very small, then so is the change in the mixing matrix and in the limit of zero noise, the EM algorithm becomes infinitely slow. Further analysis demonstrates that the first order correction term for the mixing matrix is proportional to the noiseless InfoMax update [2]. This renders the use of EM for the noise model even less attractive, since only in cases where the noise is large enough to make the $O(\sigma^4)$ contribution significant, is the solution of the noise model different from the noiseless InfoMax model. A recent result in [21] shows that going to in to hierarchical variational Bayesian framework [1] does not solve the problem. The resulting Variational Bayes EM algorithm, is suffering from the exact same defect.

With this defect of the EM algorithm in mind, we present two optimization alternatives which are closely related to the original EM algorithm.

3.2 Easy Gradient Recipe

The very appealing property of the EM algorithm is the combination of the easily implementable scheme and a guaranteed increase of the likelihood. Often, more advanced optimization methods is more demanding either analytically, computationally or both, and thus not appealing for large class of complicated problems. But using the Easy Gradient Recipe [14], one can obtain the efficiency of state-of-the-art gradient based optimization methods for the cost of the EM algorithm.This is done by recycling the E and M-steps: Consider in pseudo-code some function which is given the model parameters $\boldsymbol{\theta}$ and returns the log likelihood bound and gradient of the log likelihood bound computed in three steps

$$
\begin{aligned}
& [B, \tfrac{dB}{d\boldsymbol{\theta}}] = \texttt{fct}(\boldsymbol{\theta}) \\
\textbf{1)} \quad & \text{Find } \boldsymbol{\phi}^* \text{ such that } \left.\tfrac{\partial B}{\partial \boldsymbol{\phi}}\right|_{\boldsymbol{\phi}^*} = 0 \quad \text{(E-step)} \\
\textbf{2)} \quad & \text{Calculate } B(\boldsymbol{\theta}, \boldsymbol{\phi}^*) \\
\textbf{3)} \quad & \text{Calculate } \tfrac{\partial B(\boldsymbol{\theta}, \boldsymbol{\phi}^*)}{\partial \boldsymbol{\theta}} \quad \text{(M-step)}
\end{aligned}
$$

Step 1) is optimizing the bound with respect to the variational parameters and is therefore equivalent to the E-step. Step 2) is to compute the bound – a task which in many hidden variable problems is easy given the E-step, and step 3) is essentially the same computation as in the M-step, since the M-step is solving the stationarity condition for the model parameters. Note that in step 3) we have exploited that we in step 1) set $\boldsymbol{\phi}$ to it's value at stationarity such that implicit $\boldsymbol{\theta}$ dependence through $\boldsymbol{\phi}$ in the gradient will vanish. The returned function value and gradient can be fed to any gradient based optimizer. In

this paper we have chosen the so-called UCMINF described in [13], which is a quasi-Newton method using BFGS update, line search and trust-regions.

Note that the Easy Gradient Recipe is not merely a generalized EM algorithm. A generalized EM algorithm is increasing the bound in each step instead of maximizing it. The Easy Gradient Recipe is also a generalized EM, but is more than that: It is making it possible to approximate the log likelihood directly and maximize it with any standard gradient based method.

3.3 Overrelaxed Adaptive EM

The Easy Gradient Recipe is a very efficient approach for a modest number of parameters to be estimated, but when in the case of for example very overcomplete systems, length($\mathbf{x}$)$\ll$length($\mathbf{s}$), such as images, the number of parameters becomes too large for practical optimization. To deal with these situations, we need to introduce a third optimization approach which in some sense is a compromise between the two already presented. Among the variants of the EM algorithm that modifies the step length we find the so-called Overrelaxed Adaptive EM algorithm, which, in the M-step, takes a larger step in the direction proposed by EM,

$$\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^n + \eta(\boldsymbol{\theta}_{EM} - \boldsymbol{\theta}^n)$$

where $\eta \geq 1$. For $\eta = 1$ we retain the ordinary EM algorithm, but for each time we take a successful step forward, the parameter η is increased by a factor above 1 e.g. 2. If the bound is decreasing in a certain step, we undo the step and reset $\eta = 1$. This speeds up the process significantly and a nice feature about the Adaptive Overrelaxed EM is that the computational time spent in each step is reasonable for also a large number of parameters.

3.4 Dealing with Constrained Parameters

Some of the parameters $\boldsymbol{\theta} = \{\mathbf{A}, \boldsymbol{\Sigma}, \boldsymbol{\nu}\}$ are either by definition or application constrained to be positive, positive definite, etc. Within the hierarchical Bayesian framework this is dealt with by imposing priors on these and restricting these priors to be zero in the forbidden domains as is the case for the hidden sources $\mathbf{S}$. In the empirical Bayes setting, however, constraints can be implemented using a reparametrization and thereby avoid the trouble of the integrals otherwise involved.

The mixing matrix $\mathbf{A}$ is considered to be either unconstrained or with positive elements. The positivity constraint is constructed using the exponential func-

tion, $(\mathbf{A})_{ij} = e^{(\boldsymbol{\alpha})_{ij}}$. In this case, the parameter $\mathbf{A}$ is essentially substituted by the underlying parameter $\boldsymbol{\alpha}$ in the parameter set $\boldsymbol{\theta}$. Note that with this parametrization it holds for any function B that

$$\frac{dB}{d\left[\boldsymbol{\alpha}\right]_{ij}} = \frac{\partial \mathbf{A}}{\partial\left[\boldsymbol{\alpha}\right]_{ij}} \frac{\partial B}{\partial \mathbf{A}} = \left[\mathbf{A}\right]_{ij} \left[\frac{\partial B}{\partial \mathbf{A}}\right]_{ij} .$$

Setting this derivative to zero can be solved by a simple iterative scheme [6,9,22] as long as both the data and the sources are positive:

$$\left[\mathbf{A}\right]_{ij} := \left[\mathbf{A}\right]_{ij} \frac{\left[\boldsymbol{\Sigma}^{-1}\mathbf{X}\langle\mathbf{S}\rangle_q^T\right]_{ij}}{\left[\boldsymbol{\Sigma}^{-1}\mathbf{A}\langle\mathbf{S}\mathbf{S}^T\rangle_q\right]_{ij}} .$$

When negative data/sources are encountered the problem has to be solved via somewhat slower quadratic programming techniques.

The noise covariance must be positive definite in order to serve as a covariance, but can be further constrained to be for example diagonal. We consider in this setup noise covariances of the simple isotropic form $\boldsymbol{\Sigma} = e^{\beta}\mathbf{I}$, the diagonal form $\boldsymbol{\Sigma} = \mathrm{diag}(e^{\beta_1}, ..., e^{\beta_m})$, and the full parametrization $\boldsymbol{\Sigma} = \boldsymbol{\beta}\boldsymbol{\beta}^T$. The parameters in the source prior may also be constrained, and similar reparametrizations can be implemented.

4 Estimating Source Statistics

Calculating the required statistics and the normalization of the source posterior will in most cases be intractable because the non-Gaussian source prior and multivariate Gaussian Likelihood makes the posterior non-Gaussian multivariate. In this context tractable means we can calculate the normalization constant of the posterior, the marginal Likelihood, $p(\mathbf{X}|\boldsymbol{\theta})$ and posterior moments exactly in polynomial time. In the intractable case, we therefore have to resort to approximate inference techniques. In this section we discuss two deterministic mean field approaches, expectation consistent (EC) and variational (Bayes). In both these approaches variational distributions are used to make the approximations to the marginal Likelihood tractable, but there is an important distinction to be made between the two. In the variational approach a restricted form of the variational distribution q is used to made the calculation of the bound $B(\boldsymbol{\theta}, \boldsymbol{\phi})$ tractable. In EC, on the other hand, the aim is only for an approximation $A(\boldsymbol{\theta}, \boldsymbol{\phi})$ to the log of the marginal likelihood, which will typically be more precise than the bound. But as will be shown below in optimization of the parameters $A(\boldsymbol{\theta}, \boldsymbol{\phi})$ is used in exactly the same way as $B(\boldsymbol{\theta}, \boldsymbol{\phi})$.

4.1 Expectation Consistent

The basic idea behind the expectation consistent (EC) framework [17,15,16] is to use more than one variational distribution approximation to the posterior. These encode complementary aspects such as prior constraints and the Likelihood term. We will show below that the requirement of consistency between the distributions on the sufficient statistics, i.e. expectation consistency, follows very naturally when we derive the EC approximation to the marginal Likelihood. We will also give a recipe for attaining the consistency by a sequential iterative approach that alternates between updating each of the distributions.

In instantaneous ICA we can get tractability by choosing a decomposition into two distributions (here $\mathbf{s} = \mathbf{s}_t$ denotes the sources of one time instance t):

$$q(\mathbf{s}) = \frac{1}{Z_q(\boldsymbol{\lambda}_q)} p(\mathbf{s}) \exp(\boldsymbol{\lambda}_q^T \mathbf{g}(\mathbf{s})) \tag{7}$$

$$r(\mathbf{s}) = \frac{1}{Z_r(\boldsymbol{\lambda}_r)} p(\mathbf{x}|\mathbf{A}, \mathbf{s}, \boldsymbol{\Sigma}) \exp(\boldsymbol{\lambda}_r^T \mathbf{g}(\mathbf{s})) \ , \tag{8}$$

where the exponential factors are chosen to contain the first and diagonal second moment $\mathbf{g}(\mathbf{s}) = (s_1, \ldots, s_M, -\frac{s_1^2}{2}, \ldots, -\frac{s_M^2}{2})$, the parameters are denoted by $\boldsymbol{\lambda} = (\gamma_1, \ldots, \gamma_M, \Lambda_1, \ldots, \Lambda_M)$. Both q and r have distinct vectors $\boldsymbol{\lambda}_q$ and $\boldsymbol{\lambda}_r$ containing these terms. The normalizers are

$$Z_q(\boldsymbol{\lambda}_q) = \int d\mathbf{s}\, p(\mathbf{s}) \exp(\boldsymbol{\lambda}_q^T \mathbf{g}(\mathbf{s})) \tag{9}$$

$$Z_r(\boldsymbol{\lambda}_r) = \int d\mathbf{s}\, p(\mathbf{x}|\mathbf{A}, \mathbf{s}, \boldsymbol{\Sigma}) \exp(\boldsymbol{\lambda}_r^T \mathbf{g}(\mathbf{s})) \ . \tag{10}$$

The purpose of the exponential factors in the approximate distributions is to compensate for the factor we have omitted. How to choose the parameters will be explained below. We see that we get tractability with this choice since $q(\mathbf{s})$ is a product of univariate distributions and $r(\mathbf{s})$ is a multivariate Gaussian. Of course the choice of decomposition should be guided not only be tractability but also by quality of the approximation. We expect from central limit theorem (CLT) arguments that the EC approximation with this decomposition will become better the higher the number of sources with a "homogeneous" connectivity of the mixing matrix [15,16,5]. Empirically we observe that that the EC approximation is almost always more precise than the variational approximation even for quite small systems where we cannot really rely on the CLT argument. So in this case one may speculate that the difference is due to the fact EC is a more flexible approximation containing the variational as a

degenerate trivial second moments case, see below. To proceed we rewrite the exact marginal Likelihood as

$$p(\mathbf{x}|\mathbf{A},\mathbf{\Sigma}) = \int d\mathbf{s}\, p(\mathbf{x}|\mathbf{A},\mathbf{s},\mathbf{\Sigma})\, p(\mathbf{s}) = \frac{Z_q(\boldsymbol{\lambda}_q)}{Z_q(\boldsymbol{\lambda}_q)} \int d\mathbf{s}\, p(\mathbf{x}|\mathbf{A},\mathbf{s},\mathbf{\Sigma})\, p(\mathbf{s})$$
$$= Z_q(\boldsymbol{\lambda}_q) \Big\langle p(\mathbf{x}|\mathbf{A},\mathbf{s},\mathbf{\Sigma}) \exp(-\boldsymbol{\lambda}_q^T \mathbf{g}(\mathbf{s})) \Big\rangle_q \;, \tag{11}$$

where

$$\langle \ldots \rangle_q = \frac{1}{Z_q(\boldsymbol{\lambda}_q)} \int d\mathbf{s}\, \ldots\, p(\mathbf{s})\, \exp(\boldsymbol{\lambda}_q^T \mathbf{g}(\mathbf{s})) \tag{12}$$

denotes an average over $q(\mathbf{s})$. The first step in the EC approximation is to introduce a simpler distribution containing only the exponential factor (a product of univariate Gaussians in this case)

$$u(\mathbf{s}) = \frac{1}{Z_u(\boldsymbol{\lambda}_u)} \exp(\boldsymbol{\lambda}_u^T \mathbf{g}(\mathbf{s})) \tag{13}$$

and exchange the average over q with an average over u. If u shares some key properties with q, e.g. the two first moments, then in many cases the finer details of the distribution will not change the value of the average very much:

$$\Big\langle p(\mathbf{x}|\mathbf{A},\mathbf{s},\mathbf{\Sigma}) \exp(-\boldsymbol{\lambda}_q^T \mathbf{g}(\mathbf{s})) \Big\rangle_q \approx \Big\langle p(\mathbf{x}|\mathbf{A},\mathbf{s},\mathbf{\Sigma}) \exp(-\boldsymbol{\lambda}_q^T \mathbf{g}(\mathbf{s})) \Big\rangle_u = \frac{Z_r(\boldsymbol{\lambda}_u - \boldsymbol{\lambda}_q)}{Z_u(\boldsymbol{\lambda}_u)} \;.$$

Inserting the approximation we arrive at the EC approximation

$$A(\boldsymbol{\theta},\boldsymbol{\phi}) = \ln Z_{\mathrm{EC}}(\boldsymbol{\lambda}_q,\boldsymbol{\lambda}_u) \equiv \ln Z_q(\boldsymbol{\lambda}_q) + \ln Z_r(\boldsymbol{\lambda}_u - \boldsymbol{\lambda}_q) - \ln Z_u(\boldsymbol{\lambda}_u) \tag{14}$$

with $\boldsymbol{\phi} = \{\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_u\}$. In the following we will use a different set of parameters: With a change of variables $\boldsymbol{\lambda}_r \equiv \boldsymbol{\lambda}_u - \boldsymbol{\lambda}_q$ we can also write

$$\ln Z_{\mathrm{EC}}(\boldsymbol{\lambda}_q,\boldsymbol{\lambda}_r) = \ln Z_q(\boldsymbol{\lambda}_q) + \ln Z_r(\boldsymbol{\lambda}_r) - \ln Z_u(\boldsymbol{\lambda}_q + \boldsymbol{\lambda}_r) \tag{15}$$

and $\boldsymbol{\phi} = \{\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_r\}$. The second step in the EC approximation is to determine the parameters from the stationarity condition [17] which gives the expectation consistent condition of the three distribution

$$\frac{\partial \ln Z_{\mathrm{EC}}}{\partial \boldsymbol{\lambda}_q} = 0 : \langle \mathbf{g}(\mathbf{s}) \rangle_q = \langle \mathbf{g}(\mathbf{s}) \rangle_u \tag{16}$$
$$\frac{\partial \ln Z_{\mathrm{EC}}}{\partial \boldsymbol{\lambda}_r} = 0 : \langle \mathbf{g}(\mathbf{s}) \rangle_r = \langle \mathbf{g}(\mathbf{s}) \rangle_u \tag{17}$$

with $\boldsymbol{\lambda}_u = \boldsymbol{\lambda}_q + \boldsymbol{\lambda}_r$. At this stationarity point we have the EC approximation

$$\ln p(\mathbf{x}|\mathbf{A},\mathbf{\Sigma}) \approx A(\boldsymbol{\theta},\boldsymbol{\phi}) \;.$$

Below we will test this empirically by comparing the predictions for moments and inference in the ICA model.

EP Message Passing

Before giving explicit expressions for the marginal Likelihood expression and parameter derivatives for the ICA model, we give a general recipe for attaining the expectation consistent fixed-point which is identical to Minka's expectation propagation (EP) for two approximating factors [10]. This algorithm very often has very good convergence properties, but is not guaranteed to converge. Alternative guaranteed convergent so-called double loop algorithms exist [17]. The details for the ICA-model are given in the Appendix. Iteration k of the algorithm can be sketched as follows:

(1) Send messages from r to q
- Calculate parameters of $u(\mathbf{s})$: Solve for $\boldsymbol{\lambda}_u$: $\langle \mathbf{g}(\mathbf{s}) \rangle_u = \boldsymbol{\mu}_r(k-1) \equiv \langle \mathbf{g}(\mathbf{s}) \rangle_{r(k-1)}$
- Update $q(\mathbf{x})$: $\boldsymbol{\lambda}_q(k) := \boldsymbol{\lambda}_u - \boldsymbol{\lambda}_r(k-1)$

(2) Send messages from q to r
- Calculate parameters $u(\mathbf{s})$: Solve for $\boldsymbol{\lambda}_u$: $\langle \mathbf{g}(\mathbf{s}) \rangle_u = \boldsymbol{\mu}_q(k) \equiv \langle \mathbf{g}(\mathbf{s}) \rangle_{q(k)}$
- Update $r(\mathbf{s})$: $\boldsymbol{\lambda}_r(k) := \boldsymbol{\lambda}_u - \boldsymbol{\lambda}_q(k)$

$r(k)$ and $q(k)$ denote the distributions q and r computed with the parameters $\boldsymbol{\lambda}_r(k)$ and $\boldsymbol{\lambda}_q(k)$. Convergence is reached when $\boldsymbol{\mu}_r = \boldsymbol{\mu}_q$ since each parameter update ensures $\boldsymbol{\lambda}_r = \boldsymbol{\lambda}_u - \boldsymbol{\lambda}_q$.

EC for the ICA Model

In the following we give the explicit expressions for the EC marginal likelihood expression, eq. (14), moments and the derivatives of the marginal Likelihood approximation with respect to the parameters.

The moments and normalizer of the $q(\mathbf{s}) = \prod_i q_i(s_i)$, $i = 1, \ldots, M$, will depend upon the choice of prior. We denote the mean by

$$m_{q,i}(\gamma, \Lambda) = \frac{1}{Z_{q,i}(\gamma, \Lambda)} \int ds_i \, s_i \, p_i(s_i) \, \exp(\gamma s_i - \frac{1}{2}\Lambda s_i^2) \tag{18}$$

and likewise for the variance $v_{q,i}(\gamma, \Lambda)$. The multivariate Gaussian r-distribution has covariance and mean

$$\boldsymbol{\chi}_r = (\boldsymbol{\Lambda}_r + \mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1} \tag{19}$$
$$\mathbf{m}_r = \boldsymbol{\chi}_r(\boldsymbol{\gamma}_r + \mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}) \tag{20}$$

and normalizer

$$\ln Z_r = \frac{d-M}{2}\ln 2\pi - \frac{1}{2}\ln\det\boldsymbol{\Sigma} + \frac{1}{2}\ln\det\boldsymbol{\chi}_r + \frac{1}{2}\mathbf{m}_r^T\boldsymbol{\chi}_r^{-1}\mathbf{m}_r - \frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} \tag{21}$$

The u distribution is a the product of the univariate normals with moments $m_{u,i} = \gamma_{u,i}/\lambda_{u,i}$ and $v_{u,i} = 1/\lambda_{u,i}$. In the propagation algorithm, above and in the Appendix, we need to solve for the parameters in terms of the moments: $\gamma_{u,i} = v_{u,i}m_{u,i}$ and $\lambda_{u,i} = 1/v_{u,i}$. Finally the contribution to the marginal Likelihood from u is given by:

$$\ln Z_u = -\frac{M}{2}\ln 2\pi + \frac{1}{2}\sum_i \ln v_{u,i} + \frac{1}{2}\sum_i \frac{m_{u,i}^2}{v_{u,i}} \ . \tag{22}$$

Next we consider the derivatives of the marginal Likelihood with respect to $\mathbf{A}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\nu}$. When expectation consistency holds then $\frac{\partial \ln Z_{\rm EC}}{\partial \boldsymbol{\lambda}_q} = \frac{\partial \ln Z_{\rm EC}}{\partial \boldsymbol{\lambda}_r} = 0$ and we only need to consider the explicit parameter dependence. All $\mathbf{A}$ and $\boldsymbol{\Sigma}$ dependence is contained in $\ln Z_r$, eq. (10), and all $\boldsymbol{\nu}$ dependence in $\ln Z_q$, eq. (9). Stacking the variables, the result—which is is most easily derived by considering $\ln Z_r$ as a moment generating function—is very close to eqs. (2) and (3). Compared to these expressions the only difference is that we should now take the average with respect to the r-distribution. Note that although q and r have the same diagonal second moments, they differ on the off-diagonal terms: q has zero covariance since it factorized and r has the Gaussian covariance, eq. (19). It thus makes an important difference what variational distribution we use when calculate the derivatives. Finally, the derivatives with respect to parameters of the prior will be given by eq. (4) with the average being over the q-distribution.

4.2 Variational

The variational approximation can be motivated by the need to find a tractable expression for the bound function eq. (1). This can be achieved choosing the variational distribution in a tractable family. We have basically two possibilities for the ICA model either fully factorized $q(\mathbf{S}) = \prod_{it} q_{it}(s_{it})$ or as a multivariate Gaussian. The latter choice only give tractable expression for the variational bound for some choices of the prior. Here we will consider the fully factorized.

Choosing the variational distribution to be completely factorized it is not likely

that a perfect fit to the true source posterior is possible, but in many cases, the approximation will suffice for a successful estimation. We obtain the optimal q (in the factorized family) by setting the functional derivative $\delta B/\delta q_{it}$ equal to zero (the so-called freeform derivation) [8]. The general and specific solutions are:

$$\begin{aligned} q_{it}(s_{it}|\boldsymbol{\phi}_{it}) &= \frac{1}{c}\exp\left[\langle \ln p(\mathbf{X},\mathbf{S}|\boldsymbol{\theta})\rangle_{q\setminus q_{it}}\right] \\ &= \frac{1}{Z_q}p_i(s_{it}|\boldsymbol{\nu}_i)\exp\left[-\tfrac{1}{2}\Lambda_i s_{it}^2 + \gamma_{it}s_{it}\right] \end{aligned} \tag{23}$$

where $\langle \ldots \rangle_{q\setminus q_{it}} = \int \prod_{i't' \neq it} ds_{i't'} q_{i't'}(s_{i't'}) \ldots$ denotes an average over the variational distribution excluding $q_{it}(s_{it})$ and $\boldsymbol{\Lambda}$ (a vector of length M) and $\boldsymbol{\gamma}$ (a $M \times N$ dimensional matrix) are defined by

$$\boldsymbol{\Lambda} = \mathrm{diag}(\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A}) \tag{24}$$

$$\boldsymbol{\gamma} = \mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{X} - (\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A} - \mathrm{diag}(\boldsymbol{\Lambda}))\langle \mathbf{S}\rangle_q \ . \tag{25}$$

Note how this elegantly both provide us with the structural form of q by eq. (23) and the optimal values of the parametrization by equations for $\boldsymbol{\Lambda}$ and $\boldsymbol{\gamma}$. Note also, however, that the expression for $\boldsymbol{\gamma}$ depends on the variational mean value and the equations therefore not are closed. Using eq. (23) as a sequential update for $q(\mathbf{S})$ is the coordinate ascent algorithm for the factorized variational distribution and thus guaranteed to converge to a (local) optimum. The sufficient statistics for the variational distribution is the means because it is the only statistic which is necessary to determine the parameter $\boldsymbol{\gamma}$ and $\boldsymbol{\Lambda}$. We thus write the update equations for the variational distribution in terms of the mean function $\langle s_{it}\rangle_q = m_{q,i}(\gamma_{it}, \Lambda_i)$, eq. (18). Any convergent integral is formally a function and it may seem that we have gained little be reformulating the problem. But for a large and relevant set of source priors, the mean function $m_{q,i}$ have a nice closed form expression. In those cases we have substituted an intractable integral with a non-linear equation which evaluates much faster and more efficiently. The function $m_{q,i}$ is described for a variety of priors in [6] including binary, uniform, exponential (positive), Laplace (bi-exponential) and Gaussian.

A consequence of using the factorized variational distribution is that we will make trivial predictions for the non-diagonal second moments: $\langle s_{it}s_{i't}\rangle_q = \langle s_{it}\rangle_q\langle s_{i't}\rangle_q$ for $i \neq i'$. These second moments are used in in the derivatives of the bound function with respect to parameters, eqs. (2) and (3). This should be contrasted to the EC and the linear response correction to the variational approximation [6]. The linear response expression for the covariance is given by eq. (19) with $\Lambda_{r,it}$ being dependent upon the solution to the variational equations: $\Lambda_{r,it} = 1/v_{q,it} - \Lambda_i$. This result can thus be seen as an intermediate

step between the completely factorized variational approach and EC.

4.3 EC and Variational Comparison

In figure 1 we compare the mean squared approximation error on the first $\langle s_{it} \rangle$ and second moments $\chi_{ii't} = \langle s_{it} s_{i't} \rangle - \langle s_{it} \rangle \langle s_{i't} \rangle$ for a range of signal to noise ratios. The two RMS error measures are defined as

$$\text{Error}_1 = \left[\frac{1}{NM} \sum_{it} (\langle S_{it} \rangle_{\text{exact}} - \langle S_{it} \rangle_{\text{app}})^2 \right]^{1/2} \tag{26}$$

$$\text{Error}_2 = \left[\frac{1}{NM^2} (\sum_{ii't} \chi_{ii't,\text{exact}} - \chi_{ii't,\text{app}})^2 \right]^{1/2} , \tag{27}$$

where M is the number of sources. The example is using artificial data from a mixture of Gaussians source prior with two components (with equal weight), with zero mean and variances $\sigma_1^2 = 1, \sigma_2^2 = 0.01$. The number of samples is $N = 2000$ and the sources are mixed with a 2×2 matrix with column vectors $[1\ 0]^T$ and $[\sqrt{2}/2\ \sqrt{2}/2]^T$. For each SNR level, defined as $SNR = \text{Tr}(\mathbf{A}\langle \mathbf{s}\mathbf{s}^T \rangle \mathbf{A}^T)/\sigma^2$, a data set with the appropriate noise level is generated and thereafter solved by the various mean field approximations as indicated. In short, Figure 1 shows that expectation consistent method (EC) is much (typically orders of magnitude) more accurate than the variational method, and that linear response (VarLR) gives a huge gain in accuracy for second moments compare to the simple factorized model (VarFct).

Figure 2 shows how the gain in accuracy of the source posterior moments is influencing the overall ICA algorithm. For the two easier challenges, there is little or no gain by using the more advanced methods, but as the difficulty increases, i.e. when the column vectors of the mixing matrix becomes more co-linear, then the overall ICA algorithm is clearly benefitting from the increased accuracy of the EC approach.

5 Software

In this section we will briefly describe the `icaMF` Matlab toolbox[1] that implements the algorithms described in this paper. The basic function call is

```
[S,A,loglikelihood,Sigma]=icaMF(X,par) ,
```

[1] The toolbox with demos are available from `http://mole.imm.dtu.dk/`.

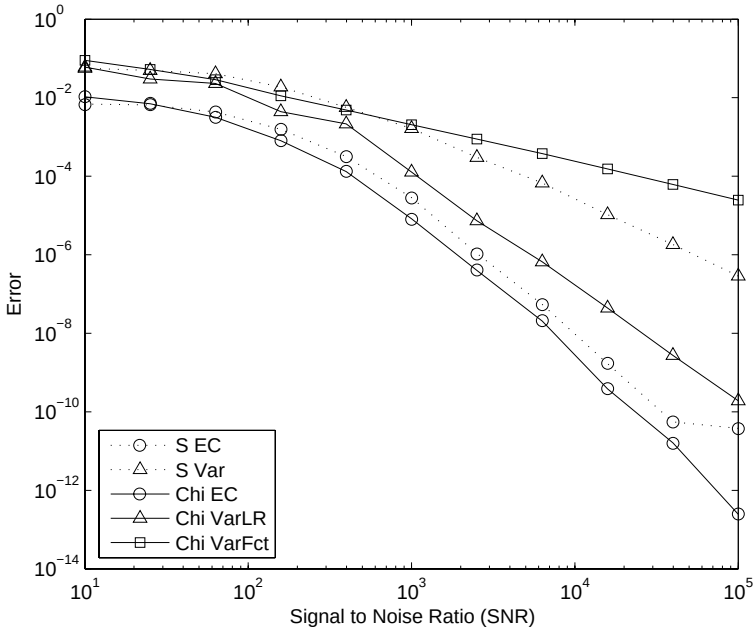

Fig. 1. Accuracy of the different posterior approximations. For a Mixture of Gaussian prior, the exact source posterior moments can be calculated and compared to the approximations. The plot show that both for the first moment (S) and the covariance (Chi), the EC is at least one order of magnitude more precise than the variational approach and the linear response (VarLR) gives an intermediate result for second moments. In [21] we also compare with the saddle-point method. It tends to give a worse approximation than variational.

where `X` is the data matrix, `par` is a list of parameters to algorithm (some of which are described below), `S` is the estimated sources, `A` is the estimated mixing matrix, `loglikelihood` is the estimated log Likelihood per sample and `Sigma` is the estimated noise covariance (default is scalar valued).

The `par` argument defines a number of settings for the algorithm. Some of the most important are summarized in table 1.

An additional feature for model selection is a Bayesian information criterion (BIC) function which calculates

$$BIC = L(\boldsymbol{\theta}) - \frac{|\boldsymbol{\theta}|}{2} \log N \ ,$$

where $|\boldsymbol{\theta}|$ is the number of parameters we estimate by maximum Likelihood, e.g. the number of free parameters in $\mathbf{A}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\nu}$. BIC is an asymptotic expansion for the log of the Likelihood marginalized over all parameter.

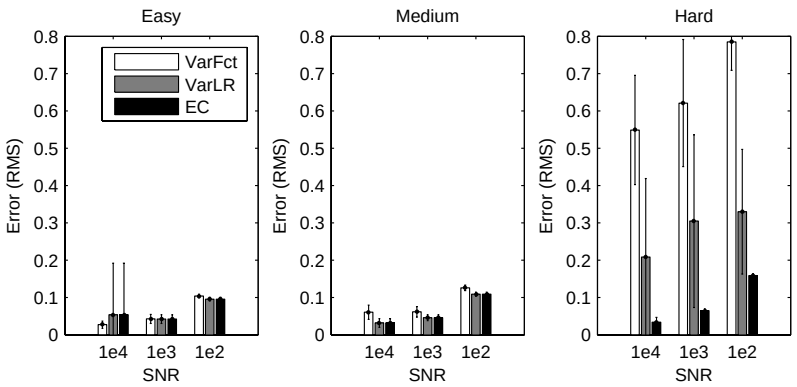

Fig. 2. Accuracy of the ICA algorithm for different posterior approximations. The three different approximations give rise to three different ICA algorithms, and the performance of these are here compared for a 2x2 case with 1000 samples and three different difficulty levels of the mixing matrix: Easy (orthogonal, $v = 3/6\pi$), medium ($v = 2/6\pi$) and hard ($v = 1/6\pi$), where v denotes the angle between the column vectors. The source prior is for both sources a zero-mean MoG with equal weights and variances $\sigma_1^2 = 0.01$, and $\sigma_2^2 = 1.99$. The error plot is average and standard deviation of the root mean square (RMS) on the estimated sources (about thirty repetitions).

par.	Usage	Default	Examples options
sources	number of sources	size(X,1)	under-/overdetermined, square
optimizer	parameter optimizer	'aem'	'em','conjgrad',bfgs'
solver	source statistics solver	'ec'	'ec','variational'
Sprior	source prior	'mog'	'exponential','binary'
method	ICA method	'free'	'constant','positive', 'fa','ppca'

Table 1
Examples of the par settings in icaMF(X,par). More detail are given in help icaMF. The ICA methods par.method explained: 'free', standard ICA, unconstrained mixing matrix and isotropic noise covariance $\mathbf{\Sigma} = \sigma^2\mathbf{I}$ optimization and heavy-tailed source prior 'mog' (fixed mixture of Gaussians); 'constant', for test sets, constant mixing matrix and noise covariance; 'positive', positive source prior 'exponential' and par.Aprior='positive'; 'fa', factor analysis; and 'ppca', probabilistic PCA.

We first tested BIC ability to find the true number of sources for an artificial data set ($N = 500$ samples, $d = 4$ dimensions and each of the $M = 3$ sources is a two-component mixture of Gaussians with with equal weight, zero means and variances 0.01 and 1.99, random mixing matrix with normal distributed entries and additive noise of variance 10^{-3}). The result, shown in Figure 3, confirms that we can find the true number of sources. Interestingly, this result

required $O(10^3)$ EM-steps showing that convergence to the maximum of the marginal Likelihood even for AOEM is slow. On the other hand the $\mathbf{A}$ matrix essentially remained unchanged after $O(10^2)$ iterations.

Figure 4 shows 1) the result of positive ICA for three sources on fMRI brain image time series (described in figure caption) and 2) the output of the function call `icaMFbic(X,par,1:5)`. For comparison, figures 5 and 6 shows the result for positive ICA and standard ICA both for four sources. Although the data has be preprocessed such that it contains both negative and positive values the result coming out of the positive ICA is more clear-cut than standard ICA. Interestingly, BIC for standard ICA tends prefer a much larger number ~ 20 of sources. This is probably because the model is more flexible than positive ICA. Note also that BIC is based upon the assumption of independent samples. This is not true in many case. In this example the samples are pixels in the image which tend to be quite correlated. So the effective number of samples are lower than the actual. Using the effective number of samples will lower the preferred number of sources but it is beyond our aim to estimate the effective number of samples.

6 Testing the Efficiency of the Framework

In this section we compare convergence properties between the optimization schemes proposed and demonstrate the strong improvement of the more advanced methods over the EM algorithm. We also shortly discuss the relation to the framework of non-negative matrix factorization, NMF.

The plots in Figure 7 are projections of the bound function contours and steps in the space of the mixing matrix. The mixing matrix is 2×2 and the space therefore 4 dimensional, but a plane is fitted to the path and bound and steps projected into this two dimensional plane. The noise is isotropic but not estimated in this example, a choice which that makes no difference to the points made. Note that the step size of the EM optimization (EM) is so small, that the dots form a solid line. In total, the EM algorithm use 729 steps to reach the optimal point. For the adaptive overrelaxed EM (AEM), the step size is far greater, and we can also see a failed test step as a detour from the region close to the optimal point. But the AEM cancels this step, resets the step size to 1 and investigate the region close to the optimal point more carefully to reach the optimal point in only 16 steps. The easy gradient approach with a quasi-Newton update (BFGS) is also doing well, reaching the optimal point in 25 steps.

Another example which links the slow convergence of the parameters with slow increase of the log likelihood can be seen in Figure 8. The data is an artificial

mix of the two speech signals available as a demo in the toolbox and the source priors are chosen to be a mixtures of Gaussians with equal weights, zero mean and variances $\sigma_1^2 = 1, \sigma_2^2 = 0.01$. As the figure shows, the development of the log likelihood approximation is rather different for the three methods. While the easy gradient approach with quasi-Newton update (BFGS) is somewhat slower in the beginning, it quickly maximizes the log likelihood to a stable higher level. As an indicator of the development of the parameters in this process, the inserts show the data and the estimated mixing matrix for the BFGS at the stable level and the EM at iteration 15 and 45. Close inspection of the inserts reveals that the EM algorithm is not at all a perfect estimation at the latter insert and this final fine-tuning takes an excessive amount of iterations.

Clearly the above demonstration is for a very specific scenario and gives no indication of the general complexity. The computational complexity of the framework as described in the following will depend crucially upon the choice of model, mean field method and optimization scheme. To recap we have N d-dimensional data points and M sources. In the E-step we will have to calculate $\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{X}$ and $\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{A}$ which is $O(NMd^2 + d^3)$ or $O(NMd)$ for full and scalar noise covariance, respectively. If the iterative solution for obtaining the moments converges in $N_{\mathrm{E-ite}}$ steps then the complexity is $O(NMN_{\mathrm{E-ite}})$, $O(NMN_{\mathrm{E-ite}} + NM^3)$ and $O(NM^3N_{\mathrm{E-ite}})$ for variational, linear response (LR) and EC, respectively. Typically the EC message passing scheme will converge in fewer steps than the variational coordinate ascent scheme somewhat compensating for the higher complexity. EM and AOEM will have the same complexity for one M-step consisting of an unconstrained $\mathbf{A}$-update $O(NM^2 + NMd + M^3)$ and $\mathbf{\Sigma}$-update $O(Nd^2)$ or $O(Nd)$ for full and scaler, respectively. The positively constrained $\mathbf{A}$-matrix update will typically be more complex because it has to be iterated eventhoug it doesn't have the M^3 term. One E-step in BFGS consists of one update (and storage) of the Hessian estimate which is quadratic in the number of parameters, $Md + d^2$ or $Md + 1$ full/scaler noise covariance, and the calculation of the gradients which is $O(d^3 + NMd + NM^2 + M^2d)$. The d^3 term disappears for scalar noise covariance. The number of iterations needed in BFGS typically will not exceed the number of parameters making the scheme scale between quadratic and cubic in the number of parameters.

In conclusion, we can handle large data sets. The complexity scales linearly with N as a simple consequence of the iid data assumption. The scaling of the complexity with input dimensionality for estimating the full noise covariance will make the ICA framework slow when d is large. For BFGS the scaling of the complexity and memory requirement with the number of parameters will not lend it impractical for large Md, as for example for images. The variational approach will be much faster than the EC and LR when M is large because of the $O(NM^3)$ complexity of the latter.

We have run extensive simulations comparing non-negative matrix factorization (NMF) [9] with positive ICA. The results for the two algorithms are very similar although in some instances where the SNR is not very high, positive ICA gives the most reasonable results. Overall NMF, like InfoMax in the unconstrained case, seems quite robust to the addition of noise. The main advantages of the ICA model are precise estimates of the noise level and the marginal likelihood at the expensive of a much slower E-step.

7 Conclusion

In this paper we have combined the improved optimization methods with the more advanced source posterior statistics and presented it in a easy-to-use toolbox. Several examples demonstrates the drawback of the traditional EM algorithm and the improved optimization obtained with adaptive overrelaxed EM and the easy gradient recipe equipped with a suitably advanced optimizer. We have also demonstrated the improved accuracy on the estimated source posterior statistics obtained by the use of the expectation consistent (EC) method as compared to the more traditional variational methods.

The upshot is an efficient ICA method which is still sufficiently flexible to encompass constraints on the source priors, mixing matrix or noise covariance. In that sense, the potential of Bayesian ICA is being realized and we believe that the toolbox implementation is presenting an easy interface to an advanced method. With a user-friendly interface, we sincerely hope that practitioners in different research disciplines will also look to the more advanced and flexible ICA methods when analyzing data in the future.

Acknowledgements

We would like to thank Lars Kai Hansen for good discussions and inform the reader that this work is funded (in part) by the Danish Technical Research Council project No. 26-04-0092 Intelligent Sound. (www.intelligentsound.org).

A Solving the EC Equations

In this appendix we give the explicit expressions for iterative scheme for solve the EC equations for the sources at time t, $\mathbf{s} = \mathbf{s}_t$:

- Initialize covariance and mean of r-distribution:

$$\boldsymbol{\chi}_r := (\boldsymbol{\Lambda}_r + \mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1} \tag{A.1}$$
$$\mathbf{m}_r := \boldsymbol{\chi}_r(\boldsymbol{\gamma}_r + \mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_t) \tag{A.2}$$

with $\boldsymbol{\gamma}_r = \mathbf{0}$ and $\boldsymbol{\Lambda}_r$ set such that the covariance is positive definite. It is sufficient to take $\boldsymbol{\Lambda}_r$ to be small positive since $\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A}$ is an outer-product form with only non-negative eigenvalues.

Run sequentially over the sources:

(1) Send message from r to q_i
- Calculate parameter of u_i: $\gamma_{u,i} := m_{r,i}/\chi_{r,ii}$ and $\Lambda_{u,i} := 1/\chi_{r,ii}$.
- Update q_i: $\gamma_{q,i} := \gamma_{u,i} - \gamma_{r,i}$ and $\Lambda_{q,i} := \Lambda_{u,i} - \Lambda_{r,i}$.
- Update moments of q_i: $m_{q,i} := m_{q,i}(\gamma_{q,i}, \Lambda_{q,i})$ and $\chi_{q,ii} = v_{q,i}(\gamma_{q,i}, \Lambda_{q,i})$.

(2) Send message from q_i to r
- Calculate parameters of u_i: $\gamma_{u,i} := m_{q,i}/\chi_{q,ii}$ and $\Lambda_{u,i} := 1/\chi_{q,ii}$.
- Update r: $\gamma_{r,i} := \gamma_{u,i} - \gamma_{q,i}$, $\Delta\Lambda_{r,i} := \Lambda_{u,i} - \Lambda_{q,i} - \Lambda_{r,i}$ and $\Lambda_{r,i} := \Lambda_{u,i} - \Lambda_{q,i}$.
- Update moments of r using Sherman-Morrison identity:

$$\boldsymbol{\chi}_r := \boldsymbol{\chi}_r - \frac{\Delta\Lambda_{r,i}}{1 + \Delta\Lambda_{r,i}\,[\boldsymbol{\chi}_r]_{ii}}[\boldsymbol{\chi}_r]_i[\boldsymbol{\chi}_r]_i^T \tag{A.3}$$
$$\mathbf{m}_r := \boldsymbol{\chi}_r(\boldsymbol{\gamma}_r + \mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_t)\ . \tag{A.4}$$

Convergence is reached when and if $\mathbf{m}_r = \mathbf{m}_q$ and $\chi_{r,ii} = \chi_{q,ii}$, $i = 1, \ldots, M$. The computational complexity of the algorithm is $O(M^3 N_{\text{ite}})$, where M is the number of sources, because each Sherman-Morrison update is $O(M^2)$ and we make M of those in each sweep over the nodes.

References

[1] H. Attias. A variational Bayesian framework for graphical models. In T. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 12*, pages 209–215. MIT Press, 2000.

[2] Anthony J. Bell and Terrence J. Sejnowski. An Information-Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[3] O. Bermond and Jean Francois Cardoso. Approximate Likelihood for Noisy Mixtures. In *Proceedings of the ICA Conference*, 1999.

[4] Mark Girolami. A variational Method for Learning Sparse and Overcomplete Representations. *Neural Computation*, 13:2517–2532, 2001.

[5] T. Heskes, M. Opper, W. Wiegerinck, O. Winther, and O. Zoeter. Approximate inference techniques with expectation constraints. *Journal of Statistical Mechanics: Theory and Experiment*, page P11015, 2005.

[6] P. Hojen-Sorensen, O. Winther, and L. K. Hansen. Mean-field approaches to independent component analysis. *Neural Computation*, 14:889–918, 2002.

[7] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10:626–634, 1999.

[8] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Mach. Learn.*, 37:183–233, 1999.

[9] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 13, pages 556–562, 2001.

[10] T. Minka. *Expectation Propagation for Approximate Bayesian Inference*. Doctoral dissertation, MIT Media Lab (2001), 2001.

[11] J. W. Miskin and D. MacKay. Ensemble learning for blind source separation. In S. Roberts and R. Everson, editors, *Independent Component Analysis: Principles and Practice*. Cambridge University Press, 2001.

[12] Radford M. Neal and Geoffrey Hinton. A new view of the EM algorithm that justifies incremental and other variants. Technical report, Department of Computer Science, University of Toronto, 1993.

[13] Hans Bruun Nielsen. UCMINF - An Algorithm for Unconstrained Nonlinear Optimization. Technical Report IMM-Rep-2000-19, Technical University of Denmark, 2000.

[14] Rasmus Kongsgaard Olsson, Tue Lehn-Schiler, and Kaare Brandt Petersen. State-space models - from the EM algorithm to a gradient approach. *Submitted to Neural Computation*, 2005.

[15] M. Opper and O. Winther. Gaussian processes for classification: Mean field algorithms. *Neural Computation*, 12:2655–2684, 2000.

[16] M. Opper and O. Winther. Adaptive and self-averaging Thouless-Anderson-Palmer mean field theory for probabilistic modeling. *Phys. Rev. E*, 64:056131, 2001.

[17] M. Opper and O. Winther. Expectation consistent approximate inference. *Journal of Machine Learning Research*, 2005.

[18] Kaare Brandt Petersen. *Mean Field ICA*. PhD thesis, Technical University of Denmark, 2005.

[19] Kaare Brandt Petersen and Ole Winther. Explaining slow convergence of EM in low noise linear mixtures. Technical Report 2005-2, Informatics and Mathematical Modelling, Technical University of Denmark, 2005.

[20] Kaare Brandt Petersen and Ole Winther. The EM Algorithm in Independent Component Analysis. In *International Conference on Acoustics, Speech, and Signal Processing*, 2005.

[21] Kaare Brandt Petersen, Ole Winther, and Lars Kai Hansen. On the convergence of EM and VBEM. *Neural Computation*, 17(9), 2005.

[22] R. Salakhutdinov and S. Roweis. Adaptive overrelaxed bound optimization methods. In *Proceedings of International Conference on Machine Learning, ICML*. International Conference on Machine Learning, ICML, 2003.

[23] H. Valpola. *Bayesian Ensemble Learning for Nonlinear Factor Analysis*. PhD thesis, Helsinki University of Technology, Espoo, 2000.

[24] W. Vanduffel, D. Fize, J. B. Mandeville, K. Nelissen, P. Van Hecke, B. R. Rosen, R. B. Tootell, and G. A. Orban. Visual motion processing investigated using contrast agent-enhanced fmri in awake behaving monkeys. *Neuron*, 32:565–577, 2001.

[25] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 13, pages 689–695, 2000.

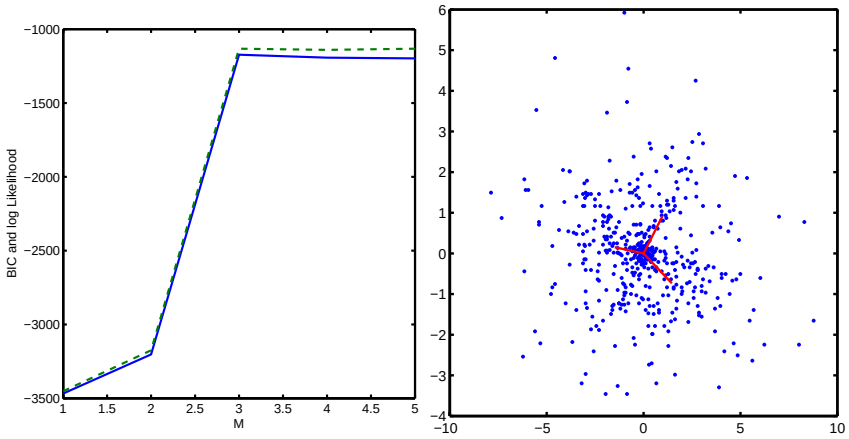

Fig. 3. Test of BIC on artificial data. The left plot shows the BIC score (full line) and the marginal log Likelihood (dashed line). The maximum for BIC is located at the true value $M = 3$ sources. The right plot shows the directions in $\mathbf{A}$-matrix and a scatter plot of the data projected to the first two data dimensions.

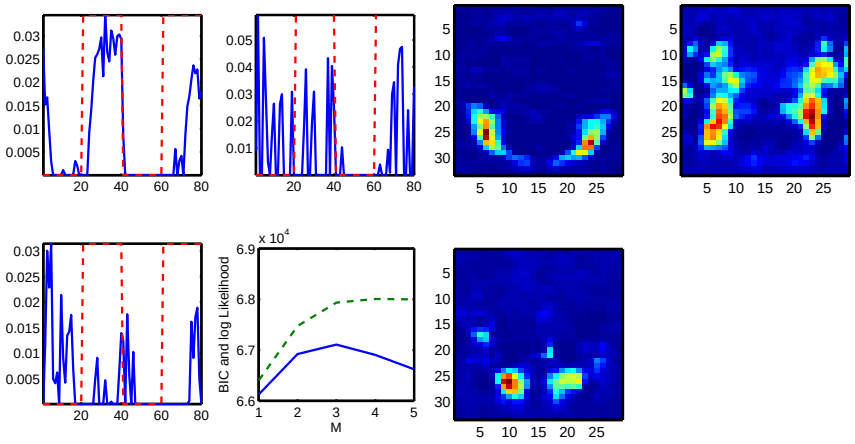

Fig. 4. "Spatial" positive ICA on fMRI time-series $\mathbf{X}$ = Time $\times$ Image for three sources. The data is described in more detail in [24]. The sub-plot with line plots shows BIC (lower full line) and the marginal log likelihood (upper dashed line) for varying number of sources. The left plot is columns in mixing matrix: time-series with visual activation paradigm rest-activation-rest-activation superimposed (dashed line). Right plot shows associated sources images. Note that the decomposition is sorted according to "energy" $E_i = \sum_d A_{di}^2 \sum_t \langle s_{it} \rangle^2$.

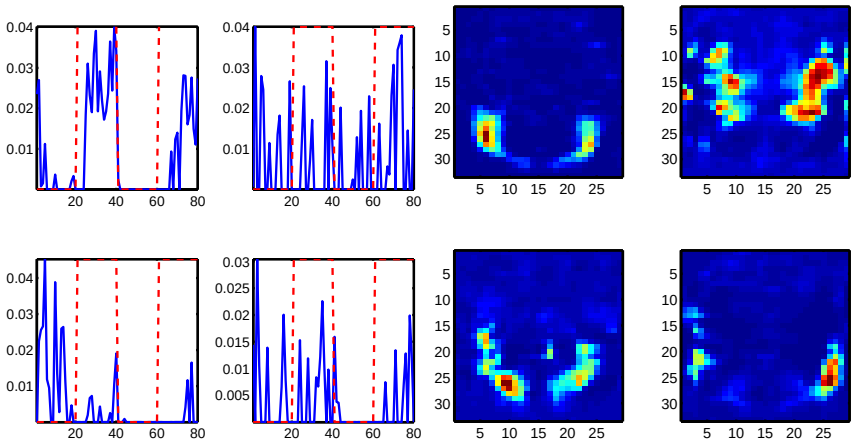

Fig. 5. "Spatial" positive ICA on fMRI time-series $\mathbf{X}$ = Time × Image for *four* sources.

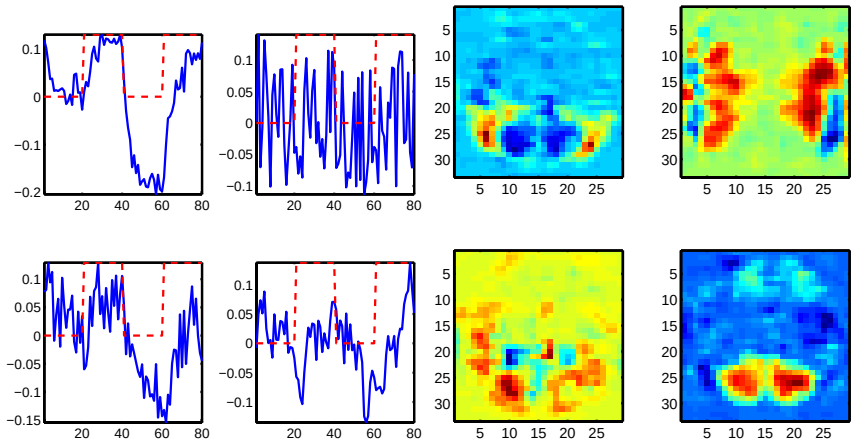

Fig. 6. "Spatial" standard ICA on fMRI time-series $\mathbf{X}$ = Time × Image for four sources. Standard ICA corresponds to unconstrained optimization of mixing matrix and heavy-tailed source prior.

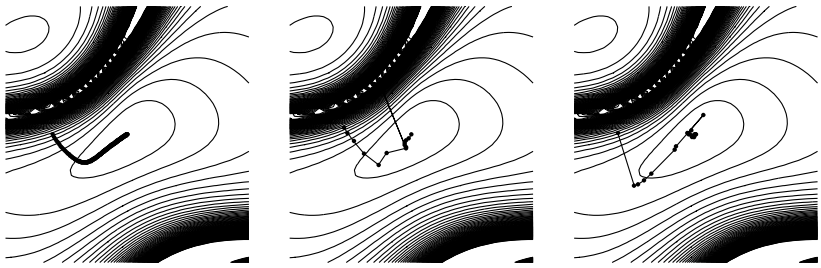

Fig. 7. Visualization of convergence. In this plot, the paths of the three different optimization methods are plotted: The EM algorithm (EM), the adaptive overrelaxed EM (AEM), and the Easy gradient with a quasi-Newton optimizer (BFGS). The contours are the bound function projected into the plane of the path.

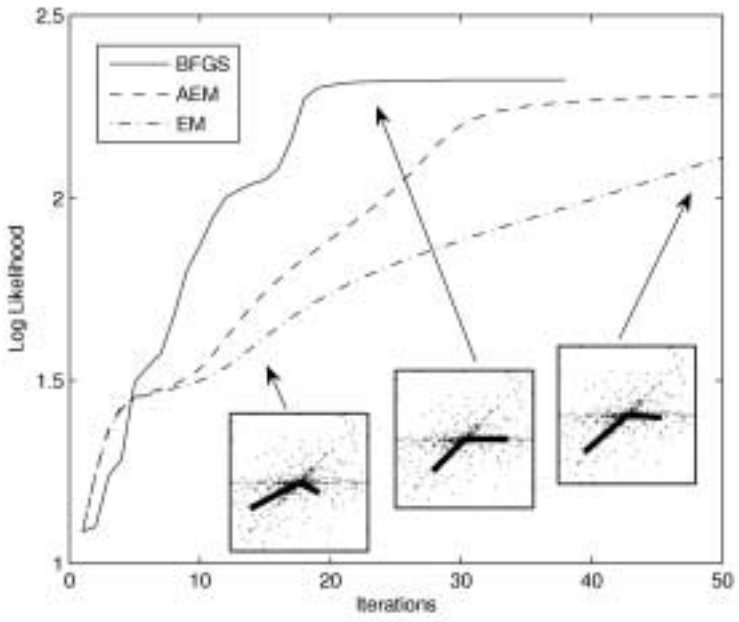

Fig. 8. Log likelihood versus iterations for the three different optimization alternatives: The EM algorithm (EM), the overrelaxed adaptive EM (AEM), and the easy gradient recipe with a quasi-Newton method (BFGS). The data is a 2×2 mixing of the speech data. The inserts shows how the estimated mixing matrix fits to the data and demonstrates that the EM algorithm is converging only slowly to the right solution compared to the BFGS method.

Authors' Answers Referee Reports NEUCOM-D-05-00490 Flexible and Efficient Implementations of Bayesian Independent Component Analysis Neurocomputing

General comments: We would like to thank the referees for a careful and constructive comments. In our revision we have followed almost all of the referees' recommendations. We give detailed answer on all points below.

Relation to previous work: The central point raised by both reviewers is about the contributions of the paper and it's relation to other work by the same authors. We feel that both 'flexible' and 'efficient' are important here. This paper can be seen as a follow-up to Ref. [6] which also describes the same flexible set-up. However, the optimization issues (efficiency) were not properly understood at the time of writing of the first paper. This seriously limited the success of the approach in some cases as also illustrated in the paper. Another major issue is the derivation of the two approximate inference methods. We think that we are now able to give a much more concise treatment of both. Adaptive TAP and expectation consistent (EC) give equivalent results when applied to the ICA model at convergence. However, the current derivation adapted from Ref. [17] give both marginals, marginal Likelihood and a message passing scheme (EP). These results and the results for optimization are also presented elsewhere [14,17,21], but we think that the ICA model in itself is so important that it merits a separate paper. Furthermore, including a software package strengthens the contribution (Neurocomputing explicitly publishes software papers).

We agree with the referees that these points needed to be clarified. We have tried our best to do this in the introduction of the paper.

Reviewer #1:

In their paper the authors propose a framework for Bayesian independent component analysis (ICA). The model is rather flexible in the choice of source prior, noise covariance structure and mixing matrix constraints. The model is learnt using either variational Bayes (VB) or expectation consistent (EC) and the authors report some comparison experiments showing that EC is preferable to VB. The problems with standard EM style updates are also discussed and the authors propose to use two alternative methods for optimizing the model parameters.

Bayesian ICA is a timely topic and the proposed framework as well as the accompanied software package are very interesting seeming to fulfill the claims of flexibility and efficiency. However, there are several issues concerning the paper that must be sorted out before it can be reconsidered for publication.

Contribution:

I had some difficulties in figuring out what is the relation of the work presented in this paper to that of ref [6]. The introduction suggests that the main

contributions are the software package and some better optimizers. But then in the main text, especially in Section 4 dealing with EC, no references to [6] are made and quite different terminology is used. So, what is the exact relation between TAP of [6] and EC of this paper?

Answer: We have clarified the relation between TAP and EC in paper.

Comparisons:

There are known problems with VB when applied to ICA which makes the comparison of Sec. 4.3 very thought provoking. However, if I understood correctly, the comparison was about inference and not learning i.e. the parameters were fixed. Since the off-diagonal posterior covariances are mostly important to what extent they make the learning more accurate and subsequently allow for better separation of the sources, it would be interesting to see a comparison in the learning setting as well, possibly with computational complexity aspects included too.

Answer: We have included a table comparing success rates for learning for different instances of easy/hard problems. We have discussed computational complexity issues in the paper.

Software:

I tried the software (a package named icaMF in the given website) and did not find it to correspond to the description of the paper. Instead of "icaMF" function I found "ica_adatap" with a somewhat different interface as described in the paper. Was it some old package the intended software not being available yet?

Answer: The new version of the software is now available at the website. Sorry for this. We wanted to wait for testing feedback before we uploaded it.

Model selection:

Since the paper advocates the usage of BIC for model order selection it would be assuring to the reader to report experiments showing that it actually works (with data where the true model is known).

Answer: We have included one such example.

I find the "model probabilities" P(M) in Figure 2 very suspicious. With these kinds of models one expects only one of the models to have a significant probability. Have you divided the log-likelihood by the number of samples or what is it that explains the unusually homogeneous distribution?

Answer: The referee is right about this point. It was an error on our part. We have consequently changed the figure to show the log-likelihood values instead.

Minor issues:

There are many small language issues all along the article. Some polishing up is necessary.

sec 3.3, line 3: "length(x) $\gg$ length(s)" "$\gg$" should be "$\ll$"

Answer: OK.

eq. 8: "$Z_q(\lambda_q)$" wrong subindices

Answer: OK.

eq. 14: a strange binary relation

Answer: OK.

eq. 18: choose whether or not to use the subindex i

Answer: OK.

fig. 1: Two figures, one for S and another for Chi would make the results much clearer

Answer: We disagree. We find the figure quite clear.

fig. 6: the x-axis should definitely be running times or flops

Answer: We think that that this figure very nicely illustrates the dynamics of learning, but of course only for a specific example. Running time or flops will be very problem specific. We have therefore instead included a paragraph on the computational complexity of the different inference and optimization methods.

Reviewer #2:

This paper describes an improved implementation of noisy ICA based on a probabilistic model. The major claims of the authors are two-fold: 1) The proposed method is efficient because two alternative optimization schemes are used to maximize the objective function in order to improve the slow convergence of EM; and 2) it is flexible with respect to prior setting in the probabilistic ICA model.

First of all, the authors should state the original contributions of this paper more clearly.

Answer: We agree and have tried to do this in the introduction.

Regarding the first point above, the use of adaptive overrelaxed EM was previously suggested in the authors' own work (Petersen & Winther, 2005), and the Easy Gradient Recipe is basically the same as the "generalized EM algorithm" (Dempster, 1977; Neal & Hinton, 1998).

Answer: Yes, but since optimization is also an important issue in the context of an ICA framework we find it in order to present it again here. Generalized EM is defined as an increase of the bound (in the formulation of Neal & Hinton 1993) or the function Q (Dempster et al. 1977), rather than a maximization. Following the Easy Gradient Recipe gives a direct approximation to maximizing the likelihood itself by any standard gradient/Newton-type method.

In addition, the flexibility of probabilistic ICA modeling, the second point, is not indeed the unique contribution of this paper. The authors can use another approximation scheme, such as Laplace approximation or sampling-based methods, even with a flexible setting of the source prior, mixing matrix, and noise covariance matrix. Furthermore, the discussion on such flexibility was also presented in a previous mean-field ICA paper (Hojen-Sorensen, et al., 2002).

Answer: Yes, it possible to perform the inference with other means and yes previous papers have also offered flexibility. But Ref. [6] didn't offer the same efficiency. We think it is in order to present work that brings together previously results when it is for an important model class like ICA.

Section 5 also obscure the focus of this paper, in which the authors describe the features of their Matlab toolbox, and also compare the positive ICA with the standard one. Both of them seems to be irrelevant.

Answer: Neurocomputing explicitly publishes software papers. Therefore we think it is in it's place to have a section on software.

My suggestion is that the authors should focus on the issue of the Expectation Consistent approximation newly introduced into the ICA context. This is interesting, and seems to be promising as shown in Figure 1.

Answer: We are happy to agree with both referees that EC applied to ICA is interesting in it's own right. We have tried to explain this in the introduction without shifting the focus too much away from the ICA framework which we also find important.

The advantage of this method in the ICA context, however, seems not to be sufficiently supported. Then, I suggest the authors to compare the performance

of EC to the previous mean-field method (including linear response and adaptive TAP) in a full ICA setting, in addition to comparing the estimated moments.

Answer: We agree. We have included additional results for learning for the different inference methods.

The optimization issue is of secondary importance, while it also contains a practical importance. So I would like the authors to present a more detailed comparison of empirical performances of the alternative EM optimizers. Figure 6 only shows a single run for each algorithm at a fixed noise level. It is better to compare the algorithms with multiple runs and different noise levels, and show clearly the advantages and drawbacks (if any) of applying these alternatives.

Answer: See answer above. The disadvantage of BFGS compared to AOEM and of EC/LR compared to variational (naive second moments) is computational complexity as also discussed in the revised version.

I also suggest the authors to reduce the irrelevant contents such as those in Section 5, make the optimization part (Section 3) more concisely, and reconsider the title of this paper such to reflect the original contribution of this paper.

Answer: Answered above. We think that the optimization part is pretty concise. This suggestion is not very specific.

Some other comments are as follows. 1) The authors' usage of the term, 'empirical Bayes' (and also 'hierarchical Bayes'), seems to be a bit strange, since the EM algorithm is referred to as 'empirical Bayes' method in this paper, while it is not a Bayesian method but a maximum-likelihood or a maximum-a-posteriori method. Please reconsider the usage or make clear the actual meaning in this paper.

Answer: We use empirical Bayes in a standard way, see e.g. Davison, Statistical Models, Cambridge, 2003. Empirical Bayes is clearly a ML method (aka MLII) because we are not averaging over all the parameters but only those that scale with the number of samples.

2) The final paragraph of Section 6 discusses the difference between their positive ICA and the Non-negative Matrix Factorization (NMF), but no simulation result is presented. Such a comparison to the standard method is useful, so I request the authors to show simulation results by NMF as well as by the previous mean-field ICA method.

Answer: We have made extensive simulations comparing NMF with positive ICA and we didn't find any differences in result that would fit well into a figure or table. This should be stated clearly in the revised paper. We have included

a comparison with linear response.