Bayesian Independent Component Analysis

Variational methods and non-negative decompositions

Winther, Ole; Petersen, Kaare Brandt

Published in:
Digital Signal Processing

Link to article, DOI:
10.1016/j.dsp.2007.01.003

Publication date:
2007

Document Version
Early version, also known as pre-print

Bayesian Independent Component Analysis: Variational Methods and Non-negative Decompositions

Ole Winther and Kaare Brandt Petersen

January 20, 2007

Abstract

In this paper we present an empirical Bayesian framework for independent component analysis. The framework provides estimates of the sources, the mixing matrix and the noise parameters, and is flexible with respect to choice of source prior and the number of sources and sensors. Inside the engine of the method are two mean field techniques – the variational Bayes and the expectation consistent framework – and the cost function relating to these methods are optimized using the adaptive over-relaxed expectation maximization (EM) algorithm and the easy gradient recipe. The entire framework, implemented in a Matlab toolbox, is demonstrated for non-negative decompositions and compared with non-negative matrix factorization.

Contents

1 Introduction

The impact of Bayesian techniques in machine learning over the last ten to twenty years can hardly be overestimated. This is largely due to the fact that the Bayesian approach is offering a structured and well-founded way of dealing with uncertainty. In machine learning, there is almost no algorithm without unknown variables and parameters to be estimated, and the Bayesian approach shows a way to handle this lack of knowledge which is intuitive and mathematically sound. One can, with the use of priors, even scale the uncertainty from complete ignorance with uniform or non-informative priors to certain knowledge with delta functions. Together with the link between priors and posteriors provided by Bayes theorem, the overall applicability of the Bayesian framework is indeed impressive and appealing.

The only downside of the Bayesian approach is the ever present integrals. These are needed for computing marginal distributions, posterior averages of the variables and marginal likelihoods. The concept of integrating out what we do not know is appealing in almost any way, for example in model selection and prediction, but is leaving us with one of the challenges which is still very hard in mathematics: To compute an integral over a general function expression. This is obviously a very general problem and over the years, researchers in many parts of science, statistics and mathematics have proposed solutions of different kinds.

In condensed matter/statistical physics, integrals equivalent to those of the Bayesian approach show up when computing properties of magnetic materials, and all through the twentieth century, many techniques and approximations have been developed. One family of techniques is the mean field methods in which the often very complicated interaction in systems with many variables is dealt with by exchanging these with a simplified (mean) influence from the other random variables. In some special cases specific mean field methods can solve the problem exactly (limit of infinite system size, Gaussian distributions, specific topologies of the problem for example trees). More importantly, in many other cases it will provide sensible approximations. Keeping a long story short, mean field techniques are approximating the very complicated integrals with fix-point solutions of non-linear equations – and for some systems, it works very well indeed. In the Bayesian context this means that the integrals can be substituted with fix-point equations of non-linear functions.

When applying a Bayesian approach to the blind source separation technique Independent Component Analysis (ICA), it is natural to integrate out at least the extensive variables, i.e. the variables which scale directly with the size of the observed data set. As presented in early work such as Refs. [2, 12], this leads to equations and parameters most easily solved by the expectation maximization (EM) algorithm. But these approaches all need to find the source statistics in some way or another. In [2], the source prior is sufficiently discrete to make exhaustive computations possible while in [12] the challenge is overcome by the use of Gaussian mixtures as source prior, which given Gaussian noise, can be integrated analytically. But these methods only work when the dimensionality

is sufficiently small and more advanced source priors are of interest and therefore techniques of finding the source statistics for these priors are relevant. One could resort to techniques like Monte Carlo sampling, but although these can provide more accurate inferences than mean field methods, it typically requires some order magnitude more computations to reach that point. So in the line of work presented here, the mean field techniques are lending themselves as a handy and efficient methodology. Some of the first ICA algorithm using mean field techniques to determine the source statistics are Refs. [11, 5], presenting good results with rather complicated source priors. The straight forward mean field approach to the problem – sometimes referred to as variational Bayes or naive mean field – is approximating the source posterior by a completely factorized distribution $q(\mathbf{s})$ for the source vector $\mathbf{s}$. Because of the factorization, the second moments are trivial products $\langle s_i s_j \rangle_q = \langle s_i \rangle_q \langle s_j \rangle_q$ for $i \neq j$, which is an obvious inadequacy. More advanced techniques such as linear response is making the covariance matrix estimate more rich [5], but is then no longer related any cost function and not consistent with the factorized model on the diagonal $\langle s_i^2 \rangle_q$ as one might wish. The expectation consistency (EC) approach is achieving both the richer structure and the consistency between the factorized distribution and covariance estimate.

As demonstrated in numerous papers, the EM algorithm is in certain cases extremely inefficient. In fact, for ICA, the EM algorithm effectively stands still in the low noise limit [3, 17]. Therefore different optimization variants have been considered such as the adaptive overrelaxed EM [19] and the easy gradient recipe with a quasi-newton optimizer [13]. These modifications have sped up the convergence with orders of magnitude compared to our original work [5].

In this paper we give a brief overview of the method with the variants of the source statistics estimation and optimization of the hyper parameters. The complete method is available as an easy-to-use Matlab toolbox and we apply the framework to non-negative decomposition of simple images (hand-written '3's [11]) comparing the Bayesian approach with non-negative matrix factorization [7, 8].

The paper is organized as follows: In section 2 we present the empirical Bayes approach to the ICA model. Section 3 describes the the two marginal likelihood approximations: the lower bound of variational Bayes and the approximation of the expectation consistent framework. In section 4 the corresponding approximations to the source statistics are derived. Section 5 sheds some light on the necessary improvements of the optimization procedure. In section 6 we consider non-negative matrix factorizations as a special case of the framework and compare with non-negative matrix factorization. Section 7 gives simulation results for decompositions of hand-written digits and section 8 is the conclusion.

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. [5]. The observation model is given by

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

with N being the number of samples and we let d denote the dimensionality of the data and M the number of sources, i.e. $\mathbf{A}$ is $d \times M$. The noise is assumed zero-mean Gaussian with covariance $\mathbf{\Sigma}$, i.e. the likelihood is $p(\mathbf{x}_t|\mathbf{A}, \mathbf{s}_t, \mathbf{\Sigma}) = \mathcal{N}(\mathbf{x}_t; \mathbf{A}\mathbf{s}_t, \mathbf{\Sigma})$. 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:

$$p(\mathbf{X}|\boldsymbol{\theta}) = \prod_t p(\mathbf{x}_t|\boldsymbol{\theta}) \tag{2}$$

$$p(\mathbf{x}_t|\boldsymbol{\theta}) = \int p(\mathbf{x}_t|\mathbf{A}, \mathbf{s}_t, \mathbf{\Sigma}) p(\mathbf{s}_t|\boldsymbol{\nu}) d\mathbf{s}_t \ . \tag{3}$$

When we are doing model selection/averaging we need to compensate for the fact that we have maximized over parameters by penalizing complex models. The Bayesian information criterion (BIC) is an asymptotic expansion for the log of the likelihood marginalized over all parameters:

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

where $\mathcal{L}(\hat{\boldsymbol{\theta}}) = \ln p(\mathbf{X}|\hat{\boldsymbol{\theta}})$ is the maximum value of the log marginal likelihood and $|\boldsymbol{\theta}|$ is the number of parameters we estimate by maximum likelihood, e.g. the number of free parameters in $\mathbf{A}$, $\mathbf{\Sigma}$ and $\boldsymbol{\nu}$. Alternatively, one may use a hierarchical Bayesian approach [1] marginalizing also over $\boldsymbol{\theta}$, see [11] for an application to ICA.

3 Approximating the Marginal Likelihood

The log marginal likelihood $\mathcal{L}(\boldsymbol{\theta}) = \ln p(\mathbf{X}|\boldsymbol{\theta})$ is for most priors of interest (e.g. heavy tailed) not tractable. Here tractable means that the computational complexity is polynomial. Besides for low dimensional cases we therefore have to resort to deterministic mean field or non-deterministic Monte Carlo methods. In this paper we will present two deterministic approaches, variational Bayes and the expectation consistent approximation.

In the celebrated *variational (Bayes)* approach [6, 1] a lower bound is used

as objective function. The lower bound $\mathcal{B}$ is defined by

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &\equiv \ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \int q(\mathbf{S}|\boldsymbol{\phi}) \frac{p(\mathbf{X},\mathbf{S}|\boldsymbol{\theta})}{q(\mathbf{S}|\boldsymbol{\phi})} d\mathbf{S} \\ &\geq \int q(\mathbf{S}|\boldsymbol{\phi}) \ln \frac{p(\mathbf{X},\mathbf{S}|\boldsymbol{\theta})}{q(\mathbf{S}|\boldsymbol{\phi})} d\mathbf{S} \equiv \mathcal{B}(\boldsymbol{\theta},\boldsymbol{\phi}) \ . \end{aligned} \tag{4}$$

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 $\mathcal{L}(\boldsymbol{\theta}) = \mathcal{B}(\boldsymbol{\theta},\boldsymbol{\phi}) + KL(q,p)$, 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. To get tractability, however, a specific *factorization* of the variational distribution is selected and then the bound, eq. (4) is maximized with respect to q in this approximating family.

In the *expectation consistent* approach [16], one is not aiming at bounding the marginal likelihood but rather approximating it, i.e. we give up the nice bounding property of the variational approximation and instead get an approximation which in general will be more precise than the bound. Here we state the results which will be derived below. For the ICA model the approximation is based upon a factorization in three terms for the prior, the likelihood and one compensating for double counting

$$\mathcal{A}(\boldsymbol{\theta},\boldsymbol{\phi}) = \sum_t \mathcal{A}_t(\boldsymbol{\theta},\boldsymbol{\phi}_t) \tag{5}$$

$$\begin{aligned}\mathcal{A}_t(\boldsymbol{\theta},\boldsymbol{\phi}_t) &= \ln \int u^*(\mathbf{s}_t;\boldsymbol{\lambda}_{q,t}) p(\mathbf{s}_t|\boldsymbol{\nu}) d\mathbf{s}_t + \ln \int u^*(\mathbf{s}_t;\boldsymbol{\lambda}_{r,t}) p(\mathbf{x}_t|\mathbf{A},\mathbf{s}_t,\boldsymbol{\Sigma}) d\mathbf{s}_t \\ &\quad - \ln \int u^*(\mathbf{s}_t;\boldsymbol{\lambda}_{q,t}+\boldsymbol{\lambda}_{r,t}) d\mathbf{s}_t \ , \end{aligned} \tag{6}$$

where

$$u^*(\mathbf{s};\boldsymbol{\lambda}) = \exp(\boldsymbol{\lambda}^T \mathbf{g}(\mathbf{s})) \tag{7}$$

is an unnormalized distribution in the exponential family: The idea here is that in each term, the parameters $\boldsymbol{\phi} = \{\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_r\}$ are set such that they compensate for the terms omitted. As we shall see in the following the parameters are derived from the stationarity condition: $\frac{\partial \mathcal{A}(\boldsymbol{\theta},\boldsymbol{\phi})}{\partial \boldsymbol{\lambda}_q} = \frac{\partial \mathcal{A}(\boldsymbol{\theta},\boldsymbol{\phi})}{\partial \boldsymbol{\lambda}_r} = 0$.

Next we will give a detailed derivation of how to estimate the approximating distribution entering in the two frameworks and after that we will make a number of general observations about how to optimize hyperparameters with empirical Bayes using not the exact log marginal likelihood but rather the bound $\mathcal{B}$ or the approximation $\mathcal{A}$.

4 Approximating Source Statistics

Since we are looking at instantaneous ICA we are dealing with N independent inference problems for the sources (given the hyperparameters). In the following

derivation we will looking at the sources for one time instance $\mathbf{s} = \mathbf{s}_t$. The joint distributions for example $q(\mathbf{S})$ is simply a product: $q(\mathbf{S}) = \prod_t q_t(\mathbf{s}_t)$.

4.1 Variational

The variational approximation can be motivated by the need to find a tractable expression for the bound function eq. (4). A popular choice that gives tractable inference for a wide range of different independent priors (e.g. discrete, mixture of Gaussians, exponential, Laplace, etc.) is a fully factorized one, $q(\mathbf{s}) = \prod_i q_i(s_i)$. An alternative is a multivariate Gaussian q, which is tractable for example for hierarchical models with the source prior (conditioned on latent variables) itself being multivariate Gaussian.

We obtain the optimal q (in the factorized family) by setting the functional derivative $\delta\mathcal{B}/\delta q_i$ equal to zero (the so-called freeform derivation) [6]. The well known general solution is

$$q_i(s_i|\boldsymbol{\phi}_i) \quad = \quad \frac{1}{c} \exp\left[\langle \ln p(\mathbf{x}_t, \mathbf{s}_t|\boldsymbol{\theta})\rangle_{q\setminus q_i}\right] \ , \tag{8}$$

where $\langle \ldots \rangle_{q\setminus q_i} = \int \prod_{i' \neq i} ds_{i'} q_{i'}(s_{i'}) \ldots$ denotes an average over the variational distribution excluding $q_i(s_i)$. The generic solution for the ICA model becomes proportional to the source prior times a univariate Gaussian:

$$q_i(s_i|\boldsymbol{\phi}_i) \quad = \quad \frac{1}{Z_q} p_i(s_i|\boldsymbol{\nu}) \exp(\boldsymbol{\lambda}_q^T \mathbf{g}(s_i)) \tag{9}$$

where the exponential form contains first and second contribution $\mathbf{g}(\mathbf{s}) = (s, -\frac{s^2}{2})$ and the parameters are denoted by $\boldsymbol{\lambda} = (\gamma, \Lambda)$. Generalizing this problem to all $M \times N$ sources we introduce $\boldsymbol{\Lambda}$ (a vector of length M) and $\boldsymbol{\gamma}$ (a $M \times N$ dimensional matrix):

$$\boldsymbol{\Lambda} \quad = \quad \mathrm{diag}(\mathbf{A}^T \boldsymbol{\Sigma}^{-1} \mathbf{A}) \tag{10}$$

$$\boldsymbol{\gamma} \quad = \quad \mathbf{A}^T \boldsymbol{\Sigma}^{-1} \mathbf{X} - (\mathbf{A}^T \boldsymbol{\Sigma}^{-1} \mathbf{A} - \mathrm{diag}(\boldsymbol{\Lambda}))\langle \mathbf{S} \rangle_q \ . \tag{11}$$

Here we have use diag to denote the (Matlab-like) operation of turning a the diagonal of matrix into a vector and turning a vector into a diagonal matrix. Note how this elegantly both provides us with the structural form of q by eq. (9) 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 are not closed. Using eq. (9) 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 are the means because they are the only statistics necessary to determine the parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\Lambda}$. We thus write the update equations for the variational distribution in terms of the mean function:

$$\langle s_{it} \rangle_q = m_{q,i}(\gamma, \Lambda) = \frac{1}{Z_{q,i}(\gamma, \Lambda)} \int ds_i \, s_i \, p_i(s_i) \, \exp(\gamma s_i - \frac{1}{2}\Lambda s_i^2) \ . \tag{12}$$

We have thus reduced the M dimensional intractable integral (or sum for discrete source priors) to a set of M non-linear equations that we can solve with guarantee plus M one-dimensional integrals which for a large and relevant set of source priors, the mean function $m_{q,i}$ have nice closed-form expressions. The function $m_{q,i}$ is described for a variety of priors in [5] including binary, uniform, exponential (non-negative), 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: $\langle s_i s_{i'} \rangle_q = \langle s_i \rangle_q \langle s_{i'} \rangle_q$ for $i \neq i'$. Although these correlations do not appear in the variational updates they will play an important and sometimes crucial role parameter estimation [20]. Fortunately, the linear response correction [5] and expectation consistent (EC) framework [16] give non-trivial estimates of correlations. There are two reasons why we in the following will concentrate on EC: There is no well defined cost function (marginal likelihood estimate) associated with linear response. Secondly, an empirical comparison in Ref. [20] shows that EC gives more precise results both in estimating source moments and in solving the ICA problem.

4.2 Expectation Consistent Framework

The basic idea behind the expectation consistent (EC) framework [16, 14, 15] is to use more than one variational distribution approximation to the posterior. None of these are on their own very precise approximations to the actual posterior but will encode complementary aspects such as prior constraints and the likelihood term. They will agree upon (be expectation consistent on) some low order statistics and be accurate in complementary ways. For example the distribution encoding the prior constraints will give good estimates for marginal distributions whereas the distribution containing the likelihood term will give precise estimates for correlations.

To be specific, for the ICA model we use the decomposition implicit in eq. (5):

$$q(\mathbf{s}) = \frac{1}{Z_q(\boldsymbol{\lambda}_q)} p(\mathbf{s}) \exp(\boldsymbol{\lambda}_q^T \mathbf{g}(\mathbf{s})) \propto p(\mathbf{s}) u^*(\mathbf{s}; \boldsymbol{\lambda}_q) \tag{13}$$

$$r(\mathbf{s}) = \frac{1}{Z_r(\boldsymbol{\lambda}_r)} p(\mathbf{x}|\mathbf{A}, \mathbf{s}, \boldsymbol{\Sigma}) \exp(\boldsymbol{\lambda}_r^T \mathbf{g}(\mathbf{s})) \propto p(\mathbf{x}|\mathbf{A}, \mathbf{s}, \boldsymbol{\Sigma}) u^*(\mathbf{s}; \boldsymbol{\lambda}_r) \ , \tag{14}$$

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 $\boldsymbol{\lambda} = (\gamma_1, \ldots, \gamma_M, \Lambda_1, \ldots, \Lambda_M)$. Note that the functional form of the q distribution is identical to the variational distribution eq. (9), whereas the multivariate Gaussian r distribution does not have a correspondence in the variational framework. This is the key to understanding the improved approximation. We have retained tractability while keeping a larger part of the correlations present in the posterior. Next we derive the approximation to the marginal distribution and conditions for the parameters of the distributions $\boldsymbol{\phi} = \{\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_r\}$. First we

make a trivial reexpression of the marginal likelihood as an average over the q distribution[1]

$$\begin{aligned} p(\mathbf{x}|\mathbf{A},\boldsymbol{\Sigma}) &= \int d\mathbf{s}\, p(\mathbf{x}|\mathbf{A},\mathbf{s},\boldsymbol{\Sigma})\, p(\mathbf{s}) = \frac{Z_q(\boldsymbol{\lambda}_q)}{Z_q(\boldsymbol{\lambda}_q)} \int d\mathbf{s}\, p(\mathbf{x}|\mathbf{A},\mathbf{s},\boldsymbol{\Sigma})\, p(\mathbf{s}) \\ &= Z_q(\boldsymbol{\lambda}_q)\Big\langle p(\mathbf{x}|\mathbf{A},\mathbf{s},\boldsymbol{\Sigma}) \exp(-\boldsymbol{\lambda}_q^T \mathbf{g}(\mathbf{s}))\Big\rangle_q \ . \end{aligned} \tag{15}$$

The key point of the approximation is then to exchange the average over q with a distribution of the exponential form, eq. (7), which has the same moments $\langle \mathbf{g}(\mathbf{S})\rangle$ as q.[2] If we make this first order cumulant expansion then in many cases the finer details of the distributions—whether we use q or u—will not change the value of the integral very much. Inserting the approximation we arrive at the EC approximation

$$\ln Z_{\mathrm{EC}}(\boldsymbol{\lambda}_q,\boldsymbol{\lambda}_u) \equiv \ln Z_q(\boldsymbol{\lambda}_q) + \ln Z_r(\boldsymbol{\lambda}_u - \boldsymbol{\lambda}_q) - \ln Z_u(\boldsymbol{\lambda}_u) \ . \tag{16}$$

With a change of variables $\boldsymbol{\lambda}_r \equiv \boldsymbol{\lambda}_u - \boldsymbol{\lambda}_q$ we get the result eq. (5).

The second step in the EC approximation is to determine the parameters from the stationarity condition [16] which gives the expectation consistent condition of the three distributions

$$\frac{\partial \mathcal{A}(\boldsymbol{\theta},\boldsymbol{\phi})}{\partial \boldsymbol{\lambda}_q} = 0 \quad : \quad \langle \mathbf{g}(\mathbf{s})\rangle_q = \langle \mathbf{g}(\mathbf{s})\rangle_u \tag{17}$$

$$\frac{\partial \mathcal{A}(\boldsymbol{\theta},\boldsymbol{\phi})}{\partial \boldsymbol{\lambda}_r} = 0 \quad : \quad \langle \mathbf{g}(\mathbf{s})\rangle_r = \langle \mathbf{g}(\mathbf{s})\rangle_u \tag{18}$$

with $\boldsymbol{\lambda}_u = \boldsymbol{\lambda}_q + \boldsymbol{\lambda}_r$. This is exactly what we expressed above: we should set the parameters such that q, r and u have the same value for the statistics on $\langle \mathbf{g}(\mathbf{s})\rangle$. Apart from this they will be very different distributions: q contains the priors and univariate Gaussian terms and r and u being multivariate and univariate Gaussian, respectively. An important remaining question is how to tune the parameters to actually fulfill the expectation consistent conditions. In [20] we give a recipe based upon Tom Minka's expectation propagation framework (EP) [9]. We also give explicit expressions for the marginal likelihood that we repeat here for completeness.

EC for the ICA Model

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 through the mean function eq. (12) and likewise we can introduce a variance function $v_{q,i}(\gamma,\Lambda)$. The multivariate Gaussian r-distribution has covariance and mean

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

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

[1] We could equivalently had used the r-distribution for this purpose.

[2] In the ICA case the u will be a product of univariate Gaussians.

and normalizer (which we need in the marginal likelihood approximation)

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

The u distribution is a the product of the univariate normals with moments $m_{u,i} = \gamma_{u,i}/\lambda_{u,i}$ and $v_{u,i} = 1/\lambda_{u,i}$. Finally the contribution to the marginal likelihood from u is given by:

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

4.3 Visualizing the Approximations

In this subsection we will try give provide some intuition about the two frameworks. What are the distributions we find approximating? Are they actually providing good approximations to the posterior? In the case of multi-modal distributions, is it necessary that the approximating distributions are themselves multi-modal in order to provide good approximations to the marginal likelihood and to marginals? It is known that the divergence measure can be tuned between in one extreme approximating the largest mode to the other extreme having an approximation which is non-zero whenever the posterior is non-zero, see [10] and references therein. The KL divergence lies in between these extremes and is typically the only tractable measure in this class of so-called α-divergences.

We have already argued that EC has an advantage compared to variational because it has two variational distributions representing complementary aspects of the exact posterior distribution. We make a visualization of this point by considering a simple two source case, where each source is drawn from a heavy-tailed distribution (two-component mixture of Gaussians with mixing proportions 0.5 and 0.5, means zero and variances 0.01 and 1, respectively). We also have two sensors, mixing matrix $\mathbf{A} = \begin{pmatrix} 1 & 1/\sqrt{2} \\ 0 & 1/\sqrt{2} \end{pmatrix}$ and isotropic observation noise with variance $\sigma^2 = 0.1$. In figure 1 we compare the posterior $p(s_1, s_2|\mathbf{x}, \mathbf{A}, \sigma^2)$, where $\mathbf{x}$ is a realized from the generative model, with $q(s_1, s_2) = q_1(s_1)q_2(s_2)$ and $r(s_1, s_2)$ for EC and $q(s_1, s_2)$ for variational. In figure 2 we compare the corresponding marginal distributions. This result is very typical. (The code to generate random examples like this can be obtained from the authors). This example illustrates the point made above. For EC, $q_i(s_i)$, $i = 1, \ldots, M$ gives quite precise approximations to the marginals (and can even handle multi-modality) and $r(\mathbf{s})$ gives a quite precise approximation to the correlations, but is too simple to catch the shape of the distribution in the high density regions. For variational, $q_i(s_i)$, $i = 1, \ldots, M$ gives a reasonable approximation to the marginals, but not as good as EC because $q(\mathbf{s})$ is a diagonal too narrow diagonal (overconfident) approximation to the posterior. Note that $q(\mathbf{s})$ in EC is wider because it is not a fit to the posterior but rather set to be consistent with r. Note also that the (Gaussian distributed) marginals of r are very poor approximations to the exact marginals.

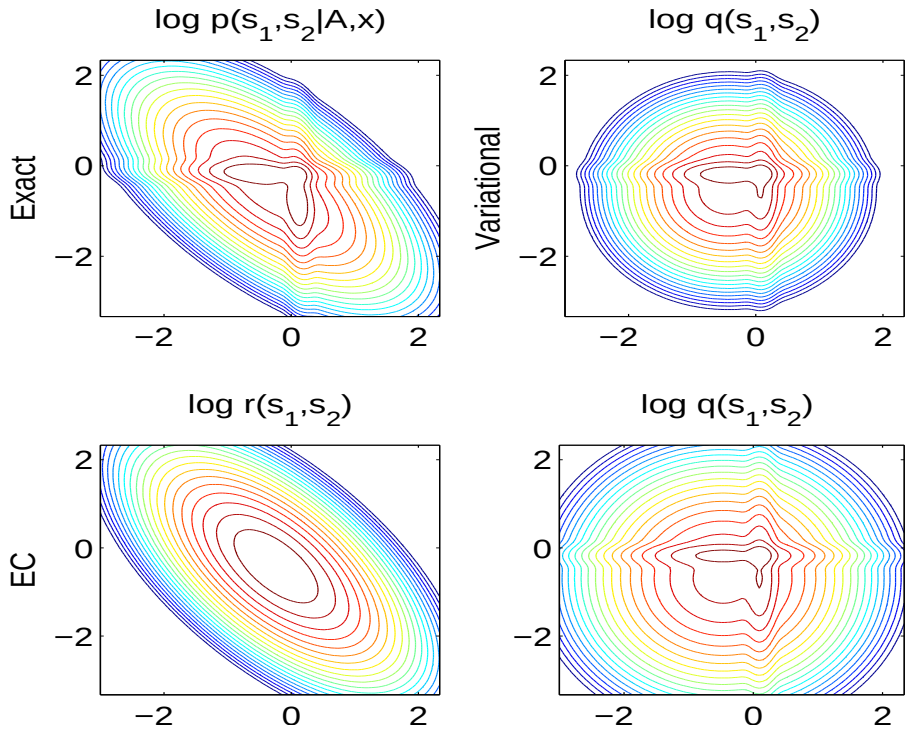

Figure 1: Contour plot of log densities for a 2d posterior distribution example with heavy-tailed prior distributions (see text for details). The upper left plot is the exact log posterior, the upper right plot is for $\log q$ for variational, the lower plots are for EC with $\log r$ left and $\log q$ right.

5 Optimization of Parameters

In this section we discuss different approaches for optimization of the marginal likelihood (approximations or exact). Although optimization of parameters is in a sense trivial because we can straightforwardly apply the EM algorithm it is of great practical importance to use schemes that are faster than the basic EM algorithm: In cases where the empirical noise estimate is small one can simply not obtain convergence in any reasonable number of iterations with the EM algorithm [3, 17]. The good news is that simple extensions—in this paper we use adaptive overrelaxed EM [19] and the easy gradient recipe [13]—are easy to implement since it is simply a matter of using the computations performed in EM in a different way. In the following we will derive the optimization techniques and then give explicit expressions for the derivatives for the ICA model for the two approximations.

For all the algorithms we need to take derivatives with respect to the pa-

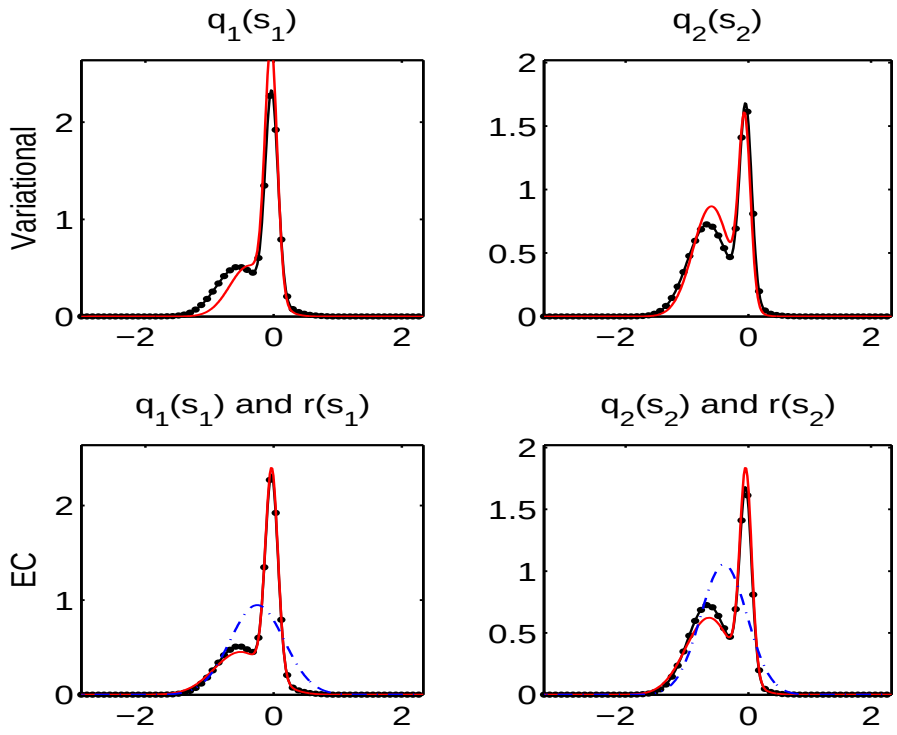

Figure 2: Marginals for the 2d example in figure 1. The two upper plots are the marginals $q_1(s_1)$ and $q_2(s_2)$ for variational (full line in red). The exact marginals are indicated by the full line with points (black). The lower plots are the same for EC with an additional dash-dotted line (blue) indicating the Gaussian distributed marginals of r.

rameters $\boldsymbol{\theta}$ (mixing matrix and noise covariance) with respect to the objective $\mathcal{B}(\boldsymbol{\theta}, \boldsymbol{\phi})$ (or $\mathcal{A}(\boldsymbol{\theta}, \boldsymbol{\phi})$). The crucial point we exploit in the algorithms is to take the derivative after solving the stationarity condition $\frac{\partial \mathcal{B}(\boldsymbol{\theta}, \boldsymbol{\phi})}{\partial \boldsymbol{\phi}} = 0$ with respect to parameters of the approximating distributions $\boldsymbol{\phi}$. This namely implies that we only need to take partial derivatives:

$$\frac{d\mathcal{B}(\boldsymbol{\theta}, \boldsymbol{\phi})}{d\boldsymbol{\theta}} = \frac{\partial \mathcal{B}(\boldsymbol{\theta}, \boldsymbol{\phi})}{\partial \boldsymbol{\theta}} + \frac{\partial \mathcal{B}(\boldsymbol{\theta}, \boldsymbol{\phi})}{\partial \boldsymbol{\phi}} \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}} = \frac{\partial \mathcal{B}(\boldsymbol{\theta}, \boldsymbol{\phi})}{\partial \boldsymbol{\theta}} \ . \tag{23}$$

In the EM algorithm we perform a coordinate ascent and are thus guaranteed to attain a local maximum of the objective. Step n in the EM-algorithm reads

$$\begin{aligned} \textbf{E-step:} \qquad & \boldsymbol{\phi}^n \leftarrow \operatorname*{argmax}_{\boldsymbol{\phi}} \mathcal{B}(\boldsymbol{\theta}^{n-1}, \boldsymbol{\phi}) \\ \textbf{M-step:} \qquad & \boldsymbol{\theta}^n \leftarrow \operatorname*{argmax}_{\boldsymbol{\theta}} \mathcal{B}(\boldsymbol{\theta}, \boldsymbol{\phi}^n) \end{aligned}$$

In the adaptive overrelaxed EM (AOEM) [19] we take steps in the EM direction with an adaptive learning rate η which can be overrelaxed, that is larger than 1:

$$\boldsymbol{\theta}^{n+1} \leftarrow \boldsymbol{\theta}^n + \eta(\boldsymbol{\theta}^n_{EM} - \boldsymbol{\theta}^n) \ .$$

Whenever the objective function decreases we backtrack to the pure EM update, and set η to one in the next step. Otherwise we multiply η with a factor above one. Depending upon the problem, η will get as large as say 40 indicating that the EM step-length is much too conservative.

In the easy gradient recipe [13] we recycle the EM computations to make a general function that evaluates the gradient and the objective. The pseudo code looks like this:

$$\begin{array}{ll} \texttt{function} & [\mathcal{B}, \frac{d\mathcal{B}}{d\boldsymbol{\theta}}] = \texttt{bound}(\boldsymbol{\theta}) \\ \textbf{1)} & \text{Find } \boldsymbol{\phi}^* \text{ such that } \frac{\partial \mathcal{B}(\boldsymbol{\theta},\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\big|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*} = 0 \quad \text{(E-step)} \\ \textbf{2)} & \text{Calculate } \mathcal{B}(\boldsymbol{\theta}, \boldsymbol{\phi}^*) \\ \textbf{3)} & \text{Calculate } \frac{\partial \mathcal{B}(\boldsymbol{\theta},\boldsymbol{\phi}^*)}{\partial \boldsymbol{\theta}} \quad \text{(M-step)} \end{array}$$

This function can be given as input to any standard effective numerical optimizer such as quasi-Newton or conjugated gradient.

5.1 Derivatives

The derivatives of the bound are easily derived for the ICA model[3]

$$\frac{\partial \mathcal{B}(\boldsymbol{\theta},\boldsymbol{\phi})}{\partial \mathbf{A}} = \boldsymbol{\Sigma}^{-1}\left(\mathbf{X}\langle\mathbf{S}\rangle_q^T - \mathbf{A}\langle\mathbf{S}\mathbf{S}^T\rangle_q\right) \tag{24}$$

$$\frac{\partial \mathcal{B}(\boldsymbol{\theta},\boldsymbol{\phi})}{\partial \boldsymbol{\Sigma}} = \frac{N}{2}\boldsymbol{\Sigma}^{-1} - \frac{1}{2}\boldsymbol{\Sigma}^{-1}\langle(\mathbf{X}-\mathbf{A}\mathbf{S})(\mathbf{X}-\mathbf{A}\mathbf{S})^T\rangle_q\boldsymbol{\Sigma}^{-1} \tag{25}$$

$$\frac{\partial \mathcal{B}(\boldsymbol{\theta},\boldsymbol{\phi})}{\partial \boldsymbol{\nu}} = \langle\frac{\partial \ln p(\mathbf{s}|\boldsymbol{\nu})}{\partial \boldsymbol{\nu}}\rangle_q \ . \tag{26}$$

The derivatives with respect to $\mathcal{A}(\boldsymbol{\theta},\boldsymbol{\phi})$ gives identical results apart from the fact that we should exchange the average with respect the q distribution in the first two equations with an average over the r distribution. This make a very significant difference since the q distribution is factorized whereas r is a multivariate Gaussian. Special care has to be taken to constrained variables [20], e.g. below we discuss multiplicative updates for non-negative $\mathbf{A}$.

6 Non-negative Decompositions

In this section we describe simulation results applying the empirical Bayes ICA framework to non-negative decompositions.[4] We describe and compare in detail

[3] We ignore subtlety that $\boldsymbol{\Sigma}$ is a symmetric matrix as we get the same result treating it as a general matrix.

[4] A Matlab toolbox, described in detail in [20], is available from `http://mole.imm.dtu.dk/` that has all the features necessary to reproduce these results.

non-negative matrix factorization (NMF) and the empirical Bayes ICA framework for the case of non-negative constraints on the mixing matrix. The main messages are that both approaches give quite similar results in the simplest set-up (isotropic noise model) but that using the Bayesian approach opens the possibility to make inferences not available in non-probabilistic approaches like NMF. More specifically in the (empirical) Bayesian approach we can estimate marginal likelihoods useful for model selection and parameters besides the mixing matrix for example a full or diagonal noise covariance matrix.

In this paper we have presented different ways to approximate source statistics and optimize parameters and also discussed that we expect EC to give more precise estimates of source statistics than variational. An obvious question is how much difference in practical situations there will be between the different methods. In our companion paper [20] we have made an effort to answer this question. Although simulations can never be exhaustive they support the following conclusions: The best results are obtained using the EC framework and the advanced optimization methods. In the following simulations we used 25 EM-steps with EC in the E-step and adaptive over-relaxed EM in the M-step (because the BFGS-optimizer cannot be applied to large parameter dimensionality). Another important empirical finding of Ref. [20] is that model selection using BIC works for known ground truth although convergence of the marginal likelihood can be slow even for advanced optimization methods. It is observed that small changes in for example the noise variance estimate, which takes many iterations to accomplish, will affect the value of the marginal likelihood significantly. Finally, non-convergence of the framework is observed in rare cases. Whether it is due to local maxima or numerical instability of EC or the mean functions (which contains erf-functions) is hard to say. Usually the problem can be solved by restarting the algorithm with different random $\mathbf{A}$.

6.1 Non-negative Matrix Factorization (NMF)

Non-negative matrix factorization (NMF) [7, 8] has become a popular technique for factorizing non-negative data into non-negative matrices, i.e. $\mathbf{X} \approx \mathbf{AS}$, where all entries of the matrices $\mathbf{A}$ and $\mathbf{S}$ are non-negative. Here we summarize the least squares NMF algorithm since it is closest in spirit to our ICA additive noise generative model eq. (1):

$$\min_{\mathbf{A},\mathbf{S}} ||\mathbf{X} - \mathbf{AS}||^2 \qquad \text{subject to} \qquad A_{li} \geq 0, S_{it} \geq 0 \;\; \text{for all} \;\; l, i, t \; .$$

It can be shown that the objecive is non-increasing using the following update rules for the two matrices [8]

$$\begin{aligned} S_{it} &\leftarrow S_{it} \frac{\left[\mathbf{A}^T\mathbf{X}\right]_{it}}{\left[\mathbf{A}^T\mathbf{AS}\right]_{it}} \\ A_{lt} &\leftarrow A_{lt} \frac{\left[\mathbf{XS}^T\right]_{lt}}{\left[\mathbf{ASS}^T\right]_{lt}} \; . \end{aligned}$$

The algorithm thus alternates in an EM-like fashion between updating all elements of each of the two matrices. This extreme simplicity and convergence guarantee have undoubtedly contributed to making NMF very popular. A further important theme is sparsity of the decomposition. Figure 5 (upper plots) gives an example where the columns of $\mathbf{A}$ are sparse. The sparsity is a consequence of the fact that only additions are allowed, compare with figure 8 and can be further enhanced by regularization, that is putting a prior on the elements of the matrices. An adaptive overrelaxed version of NMF can also be constructed [19]. We will now contrast NMF with it's empirical Bayesian counterpart.

6.2 Non-negative Empirical Bayesian ICA

NMF is in structure very similar to EM. We will thus discuss non-negative empirical Bayesian ICA in relation to EM. Generalizing the results to the optimization speed-ups discussed above is relatively strightforward, for example a gradient-type methods can be applied by reparameterizing $A_{ij} = \exp \alpha_{ij}$, where the new parameters are free [20]. In standard EM, one will in the E-step calculate the sufficient statistics of $\mathbf{S}$. In our implementation we use either variational or EC with an exponential prior for the sources to enforce non-negatively of $\mathbf{S}$: $p(s) = \exp(-s)\Theta(s)$, where $\Theta(s)$ is the Heaviside step-function. However, the following discussion about the M-step is completely general.

Setting the derivatives in eq. (24) to zero and solving with respect to $\mathbf{A}$ will not solve the problem since it will not ensure non-negativity. We can enforce this by either introducing Lagrange multipliers [5] or come up with an update rule that like NMF preserves non-negativity and is non-decreasing in the objective (marginal Likelihood bound/approximation). The following update rule

$$A_{lt} \quad \leftarrow \quad A_{lt} \frac{\left[\mathbf{\Sigma}^{-1}\mathbf{X}\langle\mathbf{S}^T\rangle\right]_{lt}}{\left[\mathbf{\Sigma}^{-1}\mathbf{A}\langle\mathbf{S}\mathbf{S}^T\rangle\right]_{lt}} \ .$$

closely resembles the NMF $\mathbf{A}$-update and was in fact suggested in Ref. [5] as a recipe for satisfying the Karush-Kuhn-Tucker conditions arising in the constrained optimization problem. If $\mathbf{\Sigma}$ is either isotropic $\mathbf{\Sigma} = \sigma^2\mathbf{I}$ or diagonal the update reduces to $A_{lt} \leftarrow A_{lt} \frac{\left[\mathbf{X}\langle\mathbf{S}^T\rangle\right]_{lt}}{\left[\mathbf{A}\langle\mathbf{S}\mathbf{S}^T\rangle\right]_{lt}}$. This update has the advantage that it is independent of $\mathbf{\Sigma}$. We can thus solve the joint M-step of $\mathbf{A}$ and $\mathbf{\Sigma}$ by first solving for $\mathbf{A}$ and then plugging the solution for $\mathbf{A}$ into the update for $\mathbf{\Sigma}$ derived from eq. (25).

There are a few important differences between NMF and empirical Bayes: Source correlations are modelled in the latter. This should make a difference in at least some cases. Working in a likelihood setting and not just with a cost function we can optimize other parameters such as the noise covariance and use the marginal likelihood in combination with BIC to make model selection. An important advantage of NMF is simplicity. One EM-step in our empirical Bayes consists of an E-step solving a set of non-linear equations to get the sufficient statistics and a M-step iterating the $\mathbf{A}$-update above to convergence (and updating the noise-covariance afterwards). In least squares NMF, an additive

normal noise model is implicitly assumed with the slight modification that negative data cannot occur. In the likelihood setting on the other hand, the model assumptions are explicit and it make sense to run the algorithm on data containing negative entries. In the following we will make an empirical investigation of the two algorithms.

7 Simulations on Hand-written Digits

We revisit the data set of 500 hand-written '3's from the MNIST database first used by Miskin and MacKay [11] to demonstrate that a purely non-negative ICA decomposition (that is all elements in both $\mathbf{A}$ and $\mathbf{S}$ are constrained to be non-negative) can yield a set of localized features representing different stroke styles. We will further extend the original analysis with a comparison with the closely related technique of non-negative matrix factorization [7, 8] and look at model selection.

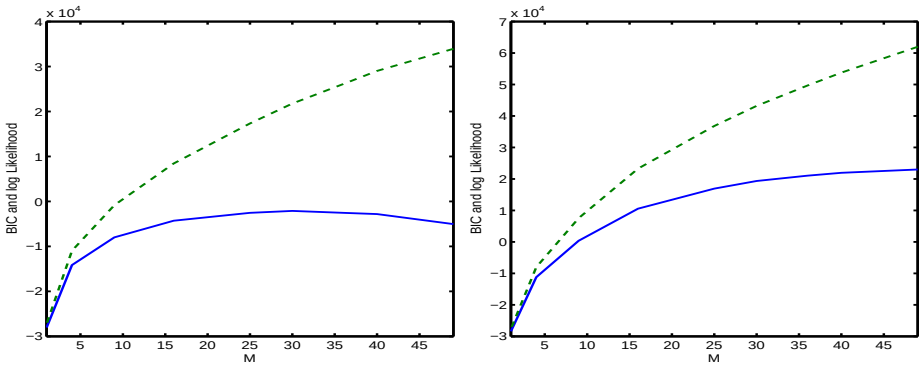

Figure 3: Marginal log likelihood (upper dashed line) and BIC score (lower full line) left for non-negative ICA and right for standard ICA as a function of the number of sources. The maximum of the BIC score is around 36 sources for non-negative ICA and at least 49 for standard ICA. Since standard ICA is a more flexible model and thus is able to fit the data more closely, it will tend to prefer larger models.

In figure 3 we compare model selection for non-negative ICA: exponential source prior and constraint optimization of $\mathbf{A}$ with more standard ICA: positive kurtosis source prior (two component equal weight zero mean Gaussian mixture with variance 1 and 0.01, respectively) and free optimization of the mixing matrix. In both cases the noise is constrained to be diagonal and iid, i.e. $\mathbf{\Sigma} = \sigma^2\mathbf{I}$, where the simple noise variance is estimated as the average of the diagonal of the full empirical Bayes noise co-variance: $\sigma^2 \leftarrow \frac{1}{d}\sum_l \Sigma_{ll}^{\text{full}}$.

Two notes of caution about the use of BIC in this context should be made: the input dimension is very large $d = 16^2 = 256$. We are therefore optimizing a

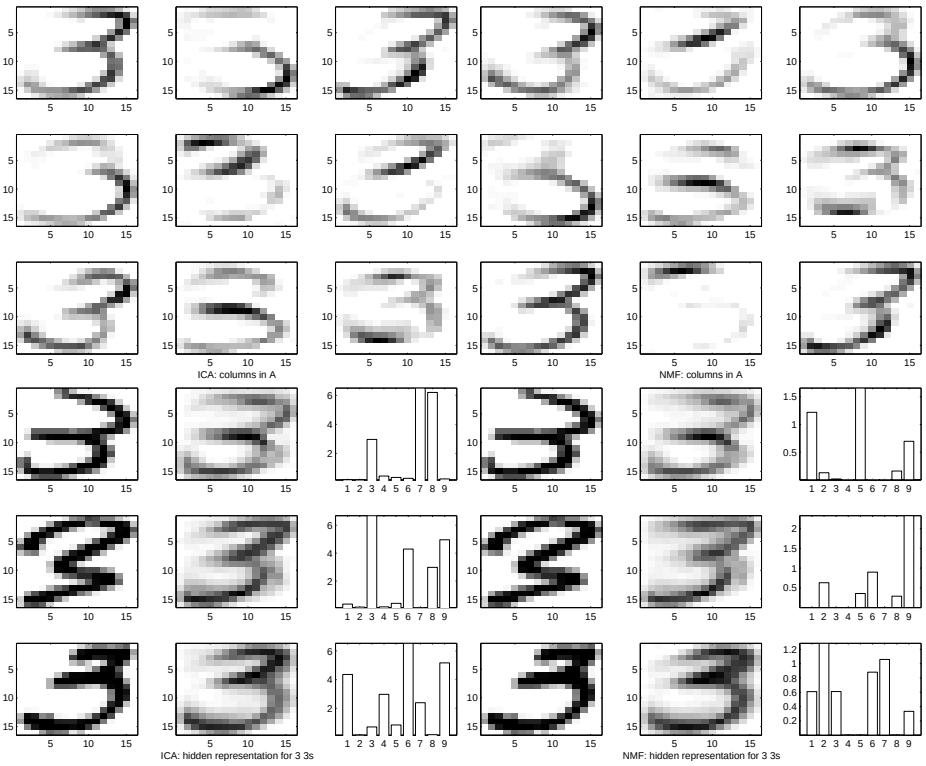

Figure 4: Upper plots: Feature images (columns in $\mathbf{A}$) for non-negative ICA (left) and NMF (right) with $M = 9$ sources. Lower left plot: From left to right three sample images, their reconstruction $\mathbf{A}\langle\mathbf{s}\rangle$ and hidden representation $\langle\mathbf{s}\rangle$ for non-negative ICA. On the right the corresponding result for NMF.

lot of parameters, $\dim(\mathbf{A})$=$d \times M$, with $d \times N$ entries in the data we might not be in the asymptotic regime required for BIC to be valid. Furthermore for non-negative ICA we are ignoring the complication dealing with constraint variables, i.e. the Laplace approximation used in BIC does not take into account non-negativity constraints. Although BIC should still give a reasonable yardstick for model selection in this case, a hierarchical Bayesian approach for example approximated by variational Bayes might be preferable [11]. It should also be noted that EC is not as readably applicable to the hierarchical Bayesian approach because when both mixing matrix and sources are both treated as random variable the r distribution, eq. (14) will no longer be tractable. Further, currently unavailable, approximations are thus needed.

From a Bayesian perspective of course if we have prior knowledge that a non-negative decomposition is more closely related to the generative process of handwriting we can disregard the standard ICA (and other type of unconstraint

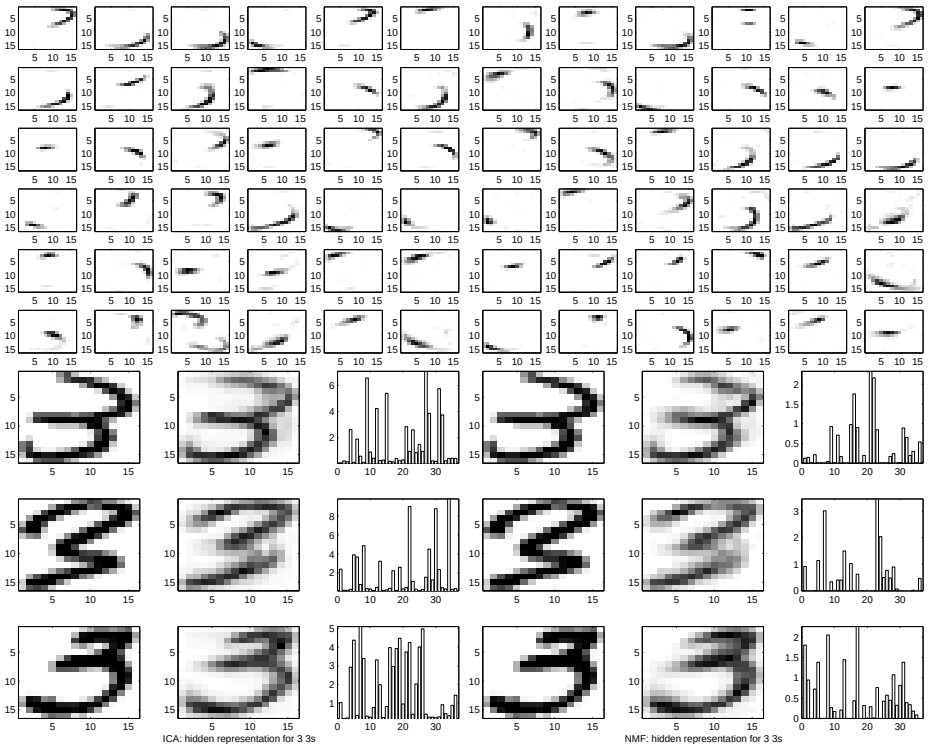

Figure 5: The same as the previous figure with $M = 36$ sources.

decomposition like PCA for that matter). However, it is still interesting to compare the results of the type as we shall do in the following.

In figures 4 and 5 we compare the feature images of non-negative ICA with non-negative matrix factorization. We see that basically the results of the two methods for very similar. Not very surprisingly they both tend to for a low number of sources to give feature images that are prototype images whereas when we increase the number of sources they become parts based. Non-negative ICA is not much slower than NMF at least up to moderate number of sources ~ 30 where the $\mathcal{O}(NM^3)$ complexity factor in EC starts to slow things down significantly.

So far we have shown simulations with a common variance for all pixels. Clearly, some pixels are almost never used and other vary a lot from image to image. Since the noise estimate is an an average over both we might believe that the less varying pixels are underfitted and more varying pixels overfitted. On the other hand the overall quality of the generative model will depend upon how we perceive the image as a whole. Therefore, it might be dangerous to weight the pixel unevenly as we do if we adapt the noise variance individually as: $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$. In figure 6 we show the result of running non-negative

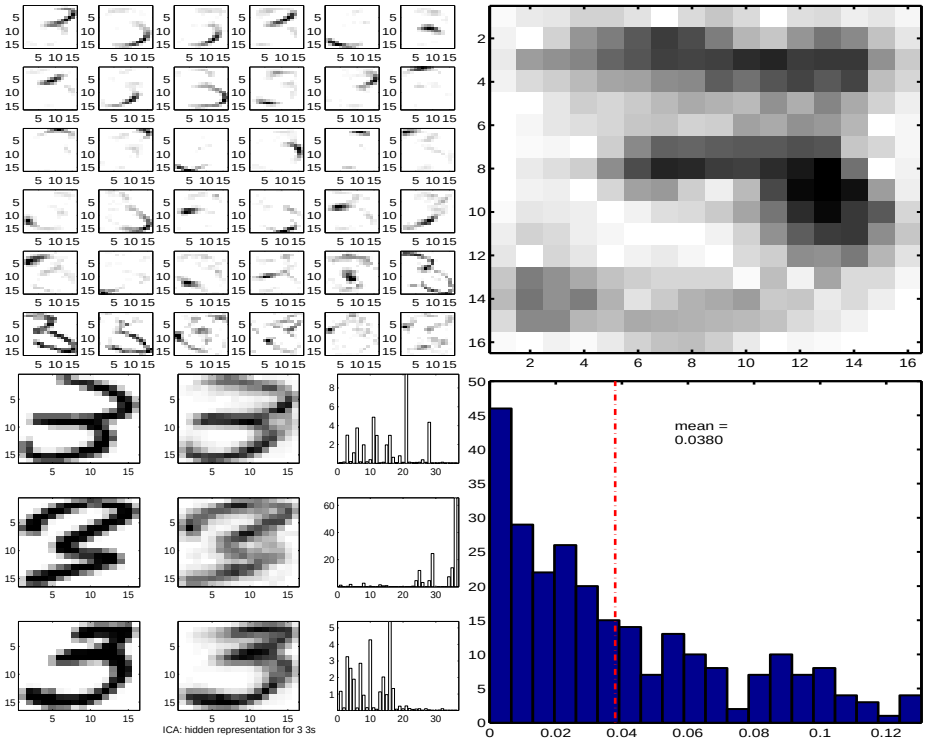

Figure 6: Non-negative ICA with diagonal noise model for $M = 36$. Upper plot left: Feature images (columns in $\mathbf{A}$). Upper right plot: noise variances represented as an image and lower right the distribution of variances. The mean value is somewhat higher than the value found for the simple model, $\sigma^2 = 0.0262$. Lower left plot: From left to right three sample images, their reconstruction $\mathbf{A}\langle\mathbf{s}\rangle$ and hidden representation $\langle\mathbf{s}\rangle$.

ICA for $M = 36$ with the diagonal noise model. We can observe a couple of things: the feature images are no longer sparse in the same way as with the simpler model. The reconstruction does not appear as nice either. This indicates that it is better to take the noise model to be simple if we want the images to look good to the human eye. Finally we can indeed see that the noise variances differ a lot across the image. In conclusion, the empirical Bayes ICA framework can fit more sophisticated noise models that the one implicit in NMF but it will be problem dependent whether they are appropriate.

Another feature of the ICA not so often used is the use on ICA on test sets, see [18] for an application to condition monitoring. The idea is simply that we tune the hyperparameters on a training set and then calculate the marginal likelihood on a test set keeping the hyperparameters fixed. This can give us a (non-Bayesian) cross-validation type test of the adequacy of the model used. In

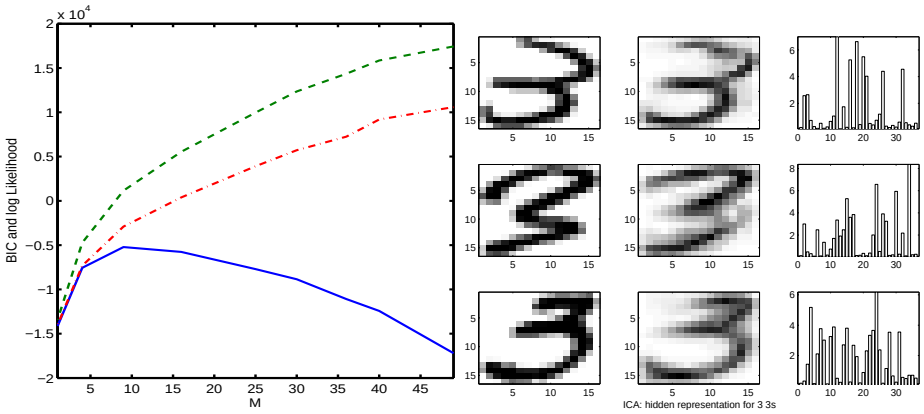

Figure 7: Running ICA on a test set. Left figure shows the marginal likelihood on the training set with 250 samples (dashed), on the test set also with 250 samples (dash-dotted) and the BIC score (full). The right plot shows the reconstruction on three test set samples for $M = 36$. These are the same samples as used above.

figure 7 we have split the training set in 250/250 in training and test. We observe that the test marginal likelihood is somewhat lower training value for all number of sources (above 1) suggesting that we are overfitting and that the model is not a perfect generative model for handwritten digits. Clearly the linear model fails to capture that some strokes are not independent, i.e. different strokes that describes the same part of the digit are mutually exclusive and some strokes will not fit together to make a digit. One can say that non-negative ICA (or NMF) can find very intuitive appealing feature images but cannot put them together. Finding models going beyond the basic linear paradigm that are capable of this grouping together of features is an important task.

Finally, we show that the standard ICA produces some reconstruction artifacts that underlines that it is a worse generative model for images than the non-negative model. Figure 8 shows the feature images and reconstruction for the same cases as figure 5.

8 Conclusion

In this paper we have described a flexible empirical Bayes framework for independent component analysis. We have used two deterministic mean field methods, variational Bayes and the expectation consistent framework, to obtain estimates of the marginal likelihood and source posterior statistics. We have presented three (hyper)parameter (mixing matrix and noise covariance) optimization schemes: the EM algorithm, adaptive overrelaxed EM and the easy gradient recipe. The EM algorithm only use gradient information whereas

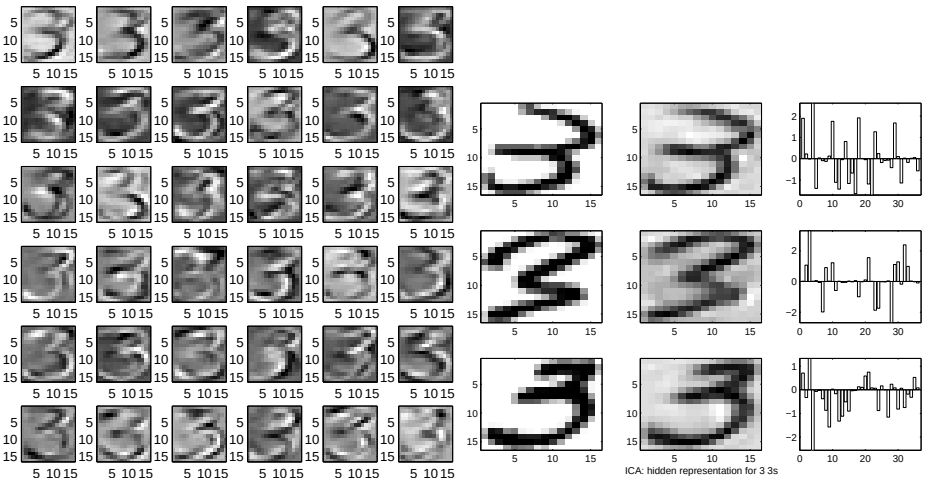

Figure 8: Standard ICA: positive kurtosis source prior and free optimization of **A** and noise variance. Should be compared with non-negative ICA figure 5. Note that the reconstruction error is smeared out over the whole image whereas the error for non-negative ICA only appears at the edges of the digit.

the latter two one need to evaluate both the gradient and the objective function (the marginal likelihood). In many practical cases it is necessary to abandon the basic EM algorithm to get convergence. Furthermore, we have made a freely available easy-to-use Matlab toolbox that supports a wide range of source priors, empirical Bayes optimization of mixing matrix and noise covariance and model selection via the Bayesian Information Criterion (BIC).

We have demonstrated the versatility of framework by finding basis function decompositions for non-negative source priors. The type of solutions found are very similar to those of non-negative matrix factorization. This is not so surprising because the two models share the same very strict constraints of allowing no subtractions. The added benefits that comes with the statistical framework are model selection and the ability to deal with other parameters. But perhaps the most interesting thing about the very appealing parts based images that comes out of linear non-negative decompositions is that the actually point to the fact that these models are not good generative models of images. Some parts, for example different stroke styles for the same part of had-written digit, are mutually exclusive and thus not well-described by a linear model. With the more complicated non-linear models that will be needed for this task (a good aproximation to) Bayesian inference will be even more important. These models could range from ICA models with coupled source priors to mixture models and layered network models [4].

Seemingly the instantaneous ICA problem have been receiving less attention the last few years from the machine learning community, but this is not due to

the problems being solved or complete flexibility is achieved. Techniques for instantaneous ICA which are fast, reliable *and* at the same time flexible with respect to choice of source prior and the number of sensors and sources, are still very rare. We hope with the presented approach based on variational mean field techniques to have contributed to this direction of research and inspired others to embrace the theoretical beauty and practical significance of this approximation to Bayesian ICA. We see the current interest in very sparse non-negative decompositions as one place where ICA with specific prior distributions can be useful.

Acknowledgements

This work is funded (in part) by the Danish Technical Research Council project No. 26-04-0092 "Intelligent Sound" (www.intelligentsound.org).

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] Adel Belouchrani and Jean-Francois Cardoso. A maximum likelihood source separation for discrete sources. In *Proceedings of EUSIPCO*, volume 2, pages 768–771, 1994.

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

[4] G. E. Hinton, S. Osindero, and Y. W. Teh. A fast learning algorithm for deep belief networks. *Neural Comput.*, 18:1527–1554, 2006.

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

[6] 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.

[7] D. Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[8] 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.

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

[10] T. P. Minka. Divergence measures and message passing. Technical report, Microsoft Tech Report, 2005.

[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] Eric Moulines, Jean-Francois Cardoso, and Elisabeth Gassiat. Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *Proceedings of the ICA Conference*, 1997.

[13] Rasmus Kongsgaard Olsson, Tue Lehn-Schiler, and Kaare Brandt Petersen. State-space models - from the EM algorithm to a gradient approach. *Neural Computation (accepted - in press)*, 2006.

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

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

[16] M. Opper and O. Winther. Expectation consistent approximate inference. *Journal of Machine Learning Research*, 6:2177–2204, 2005.

[17] Kaare Brandt Petersen, Ole Winther, and Lars Kai Hansen. On the convergence of EM and VBEM. *Neural Computation*, 17:1921–1926, 2005.

[18] N. H. Pontoppidan, S. Sigurdsson, and J. Larsen. Condition monitoring with mean field independent components analysis. *Mechanical Systems and Signal Processing*, 19(6):1337–1347, nov 2005. Special Issue: Blind Source Separation.

[19] 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.

[20] Ole Winther and Kaare Brandt Petersen. Flexible and efficient implementations of bayesian independent component analysis. *Submitted to Neurocomputing*, 2006.