Mean Field Approaches to Independent Component Analysis

Pedro A.d.F.R. Højen-Sørensen[1], Ole Winther[1,2] and Lars Kai Hansen[1]

[1]Department of Mathematical Modelling,
Technical University of Denmark, Building 321
DK-2800 Lyngby, Denmark

[2]Center for Biological Sequence Analysis
Department of Biotechnology
Technical University of Denmark, Building 208
DK-2800 Lyngby, Denmark
phs,lkhansen@imm.dtu.dk and winther@cbs.dtu.dk

January 26, 2001

Abstract

We develop mean field approaches for probabilistic independent component analysis (ICA). The sources are estimated from the mean of their posterior distribution and the mixing matrix (and noise level) is estimated by maximum a posteriori (MAP). The latter requires the computation of (a good approximation to) the correlations between sources. For this purpose we investigate three increasingly advanced mean field methods: variational, linear response and adaptive TAP and test the resulting algorithms on a number of problems. On synthetic data the advanced mean field approaches are able to recover the correct mixing matrix in cases where the variational mean field theory fails. For hand-written digits, sparse encoding is achieved using non-negative source and mixing priors. For speech, the mean field method is able to separate in the underdetermined (overcomplete) case of two sensors and three sources. One major advantage of the proposed method is its generality and implementational simplicity. Finally, we point out several possible extensions of the approaches developed here.

1 Introduction

Reconstruction of statistically independent source signals from linear mixtures is an active research field with numerous important applications, for background and references see e.g. [Lee 1998; Girolami 2000]. Blind signal separation in the face of additive noise typically involves four estimation problems: Estimation of source signals, source distribution, mixing coefficients, and noise distribution.

A full Bayesian treatment of the combined estimation problem is possible but requires extensive Monte Carlo sampling [Belouchrani and Cardoso 1995], therefore several authors have proposed variational (aka mean field or ensemble) approaches in which

the posterior distributions are either approximated by factorized Gaussians and/or integrals over the posteriors are evaluated by saddle point approximations [Attias 1999; Belouchrani and Cardoso 1995; Lewicki and Sejnowski 2000; Lappalainen and Miskin 2000; Hansen 2000; Rowe 1999; Knuth 1999]. The resulting algorithm is an Expectation-Maximization (EM) like procedure with the four estimations performed sequentially. One important problem with these approximations arises from the assumed posterior independence of sources. In particular, variational mean field theory using factorized trial distributions only treats "self-interactions" correctly, while producing trivial second moments, i.e. $\langle S_i S_i \rangle = \langle S_i \rangle \langle S_j \rangle$ for $i \neq j$. This is a poor approximation when estimating the mixing matrix and noise distribution since these estimates will typically depend upon correlations.

Recently, Kappen and Rodríguez [Kappen and Rodríguez 1998] pointed out that for Boltzmann Machines this naive mean-field (NMF) approximation — introduced in this context by [Peterson and Anderson 1987] — may fail completely in some cases. They went on to propose an efficient learning algorithm based on linear response (LR) theory. LR theory gives a recipe for computing an improved approximation to the covariances directly from the solution to the NMF equations [Parisi 1988]. In this paper, we give a general presentation of LR theory and apply it to the probabilistic ICA problem. We also briefly outline the supposedly more accurate adaptive TAP mean field theory [Opper and Winther 2000b] and compare this method to the NMF and LR approach. Whereas estimates of correlations obtained from variational mean field theory and its linear response correction in general differ, adaptive TAP is constructed such that it is consistent with linear response theory.

We expect that advanced mean field methods such as LR and TAP can be useful in the many contexts within neural computation, where variational mean field theory already have proven to be useful, e.g. for sigmoid belief networks [Saul et al. 1996]. In our experience, the main difference between variational mean field and the advanced methods lies in the estimates of correlations (often needed in algorithms of the EM-type) and the calculation of the likelihood of the data. We will not discuss the latter here, however, see [Opper and Winther 2000b] for a general method for computing the likelihood from the covariance matrix. In ICA simulations, we find that the variational approach can fail typically by ignoring some of the sources and consequently overestimating the noise covariance. The LR and TAP approaches on the other hand succeed in all cases studied. However, we do not find a significant improvement using TAP (which is also somewhat more computationally intensive), suggesting that LR is close to being the optimal mean field approach for the probabilistic ICA model.

The derivation of the mean-field equations are valid for a general source prior (without temporal correlation) and tractable for priors that can be folded analytically with a Gaussian distribution. This includes mixture of Gaussians, Laplacian and binary distributions. For other priors, one has to evaluate an extensive number of one dimensional integrals numerically. Alternatively, one can construct computationally tractable ICA algorithms using priors that are only defined implicitly. To illustrate this point we define one such algorithm which approximately corresponds to the prior having a power law tail.

To underline the flexibility and computational power of the probabilistic ICA framework and its mean field implementation, we give two quite different real world examples of recent interest that straight forwardly can be solved in this framework. The first example is that of separating speech in the overcomplete setting of two sensors and three sources [Lewicki and Sejnowski 2000] using a heavy tailed source prior such a Laplacian or the (approximative) power law prior described above. The second real world problem

considered in this paper is that of feature extraction in images. For images, it is natural to work with a non-negativity constraint for the mixing matrix and sources as in [Lee and Seung 1999]. In the probabilistic framework this type of prior knowledge is readily built into the mixing matrix and source priors.

Throughout this paper we confine ourselves to fixed source priors. There are, however, no theoretical problems in extending the EM algorithm to estimating hyperparameters, see e.g. [Attias 1999] for an example of such source prior parameter estimation.

The paper is organized as follows. In section 2 the basic probabilistic ICA model and the associated learning problem is stated. Section 3 concerns the inference part of the learning problem; we will see that variational mean field theory, linear response theory and the adaptive TAP approach can be seen as stepwise more refined ways of estimating correlations. Applying the advanced mean field methods to independent component analysis is the main contribution of this paper. Another contribution is the generality of the framework: In section 4 we examine various types of explicitly given source priors which in turn leads us to define an implicitly given source prior. The impatient or application minded reader might consult section 4.1 which shows a tabel summarizing all priors considerd in this paper. Section 5 shows some simulation results on both synthetic data and on real world data. Finally, obvious ways to extend this work is outlined in the conclusion given in section 6. The pseudo-code for the algorithm is outlined in appendix A and some additional priors not directly used in this paper are given in appendix B.

2 Probabilistic ICA

We formulate the ICA problem as follows [Hansen 2000]: The measurements are a collection of N temporal D-dimensional signals $\mathbf{X} = \{X_{dt}\}$, $d = 1, \ldots, D$ and $t = 1, \ldots, N$, where X_{dt} denotes the measurement at the dth sensor at time t. Similarly, let $\mathbf{S} = \{S_{mt}\}$, $m = 1, \ldots, M$, denote a collection of M mutually statistical independent sources, where S_{mt} is the mth source at time t. The measured signal $\mathbf{X}$ is assumed to be an instantaneous linear mixing of the sources corrupted with additive white Gaussian noise $\mathbf{\Gamma}$ that is,

$$\mathbf{X} = \mathbf{AS} + \mathbf{\Gamma} \ , \tag{1}$$

where $\mathbf{A}$ is a (time independent) mixing matrix and the noise is assumed to be without temporal correlations and with time independent covariance matrix $\mathbf{\Sigma}$, i.e. we have $\overline{\Gamma_{dt}\Gamma_{d't'}} = \delta_{tt'}\Sigma_{dd'}$. We thus have the following likelihood for parameters and sources,

$$P(\mathbf{X}|\mathbf{A}, \mathbf{\Sigma}, \mathbf{S}) = (\det 2\pi\mathbf{\Sigma})^{-\frac{N}{2}} \, e^{-\frac{1}{2}\operatorname{Tr}(\mathbf{X}-\mathbf{AS})^T\mathbf{\Sigma}^{-1}(\mathbf{X}-\mathbf{AS})} \ . \tag{2}$$

The aim of independent component analysis is to recover the unknown quantities: the sources $\mathbf{S}$, the mixing matrix $\mathbf{A}$ and the noise covariance $\mathbf{\Sigma}$ from the observed data.

The main difficulty is associated with estimation of the source signals. The estimation problems for the mixing matrix and the noise convariance matrix are relatively simple, given the sufficient source statistics. Hence, our primary objective is to improve on the estimate of sufficient statistics from the posterior distribution of the sources. The mixing matrix $\mathbf{A}$ and the noise covariance $\mathbf{\Sigma}$ are then in turn estimated by maximum a posteriori (MAP) (or maximum likelihood II (ML-II)). This naturally leads to a EM-type algorithm where the expectation step amounts to finding the posterior mean and covariances of the sources and the maximization step is the MAP/ML-II estimation. Mean field methods especially the advanced ones are well suited for the non-trivial expectation step.

Given the likelihood eq. (2), the posterior distribution of the sources is readily given by,

$$P(\mathbf{S}|\mathbf{X},\mathbf{A},\mathbf{\Sigma}) = \frac{P(\mathbf{X}|\mathbf{A},\mathbf{\Sigma},\mathbf{S})P(\mathbf{S})}{P(\mathbf{X}|\mathbf{A},\mathbf{\Sigma})} \ , \tag{3}$$

where $P(\mathbf{S})$ is a prior on the sources which might include temporal correlations (although we will postpone this problem to a future contribution [Højen-Sørensen et al. 2001]).

2.1 Estimation of mixing matrix and noise covariance

The likelihood of the parameters is given by,

$$P(\mathbf{X}|\mathbf{A},\mathbf{\Sigma}) = \int d\mathbf{S} P(\mathbf{X}|\mathbf{A},\mathbf{\Sigma},\mathbf{S})\ P(\mathbf{S}) \ . \tag{4}$$

The problem of estimating the mixing matrix and noise covariance now amounts to finding the saddle-points of the likelihood eq. (4) wrt. the mixing matrix and noise covariance. We note that the saddle-points will be given in terms of averages over the source posterior. These calculations of mean sufficient statistic wrt. to the posterior are the main challenge for mean field approaches since the sources will be coupled through the observations.

The mixing matrix $\mathbf{A}$ will be estimated by maximum a posteriori (MAP) and the noise by ML-II for convenience,

$$\mathbf{A}_{\text{MAP}} = \operatorname*{argmax}_{\mathbf{A}} P(\mathbf{A}|\mathbf{X},\mathbf{\Sigma}) \tag{5}$$

$$\mathbf{\Sigma}_{\text{MLII}} = \operatorname*{argmax}_{\mathbf{\Sigma}} P(\mathbf{X}|\mathbf{A},\mathbf{\Sigma}) \ , \tag{6}$$

where the posterior of $\mathbf{A}$ is given by $P(\mathbf{A}|\mathbf{X},\mathbf{\Sigma}) \propto P(\mathbf{X}|\mathbf{A},\mathbf{\Sigma})P(\mathbf{A})$, where $P(\mathbf{A})$ is the prior on $\mathbf{A}$. For the optimization in eqs. (5) and (6), we need the derivatives of the likelihood term,

$$\frac{\partial}{\partial \mathbf{A}} \log P(\mathbf{X}|\mathbf{A},\mathbf{\Sigma}) = \mathbf{\Sigma}^{-1}(\mathbf{X}\langle\mathbf{S}\rangle^T - \mathbf{A}\langle\mathbf{S}\mathbf{S}^T\rangle) \tag{7}$$

$$\frac{\partial}{\partial \mathbf{\Sigma}} \log P(\mathbf{X}|\mathbf{A},\mathbf{\Sigma}) = \frac{1}{2}\mathbf{\Sigma}^{-1}\langle(\mathbf{X}-\mathbf{A}\mathbf{S})(\mathbf{X}-\mathbf{A}\mathbf{S})^T\rangle\mathbf{\Sigma}^{-1} - \frac{N}{2}\mathbf{\Sigma}^{-1} \ , \tag{8}$$

where $\langle\cdot\rangle = \langle\cdot\rangle_{\mathbf{S}|\mathbf{A},\mathbf{\Sigma},\mathbf{X}}$ denotes the posterior average wrt. the sources given the mixing matrix and noise covariance. Equating eq. (8) to zero leads to the well known result for $\mathbf{\Sigma}$,

$$\mathbf{\Sigma}_{\text{MLII}} = \frac{1}{N}\langle(\mathbf{X}-\mathbf{A}\mathbf{S})(\mathbf{X}-\mathbf{A}\mathbf{S})^T\rangle \ . \tag{9}$$

In the particular case of measurements with i.i.d. noise we can simplify the covariance $\mathbf{\Sigma} = \sigma^2\mathbf{I}$, hence $\sigma^2 = \operatorname{Tr}\mathbf{\Sigma}_{\text{MLII}}/D$, where D is the number of sensors.

For $\mathbf{A}$, we consider two factorized priors, $P(\mathbf{A}) = \prod_{dm} P(A_{dm})$, a zero mean Gaussian $P(A_{dm}) \propto \exp(-\alpha_{dm}A_{dm}^2/2)$ and the Laplace distribution $P(A_{dm}) \propto \exp(-\beta_{dm}|A_{dm}|)$. Furthermore, we consider optimizing A_{dm} both unconstrained and constrained to be non-negative. Clearly, the MAP approach offers a flexibility for encoding prior knowledge about $\mathbf{A}$ that are not available in the maximum likelihood II approach, i.e. one can encode sparseness [Hyvärinen and Karthikesh 2000] and non-negativeness (for e.g. images and text, see section 5 and [Lee and Seung 1999]).

Unconstrained mixing matrices. A straight forward calculation give us the following iterative equation for the MAP estimate of $\mathbf{A}$

$$\mathbf{A}^{(k+1)} = \left(\mathbf{X}\langle\mathbf{S}\rangle^T - \mathbf{\Sigma}(\alpha\mathbf{A}^{(k)} + \beta\text{sign}(\mathbf{A}^{(k)}))\right)\langle\mathbf{S}\mathbf{S}^T\rangle^{-1} , \tag{10}$$

where we have included both priors and set $\alpha_{dm} = \alpha$ and $\beta_{dm} = \beta$. This equation can be solved explicitly for the Gaussian prior with equal noise variance on all sensors, i.e. $\beta = 0$ and $\Sigma = \sigma^2\mathbf{I}$

$$\mathbf{A} = \mathbf{X}\langle\mathbf{S}\rangle^T\left(\langle\mathbf{S}\mathbf{S}^T\rangle + \alpha\sigma^2\mathbf{I}\right)^{-1} . \tag{11}$$

The ML-II estimate is the special case obtained by setting $\alpha = 0$.

Non-negative mixing matrices. To enforce non-negative $\mathbf{A}$, we introduce a set of non-negative Lagrange multipliers $L_{dm} \geq 0$ and maximize the modified cost: $\log P(\mathbf{A}|\mathbf{X}, \mathbf{\Sigma}) + \text{Tr}\,\mathbf{L}^T\mathbf{A}$. Solving for the Lagrange multipliers we get,

$$\mathbf{L} = \mathbf{\Sigma}^{-1}(\mathbf{A}\langle\mathbf{S}\mathbf{S}^T\rangle - \mathbf{X}\langle\mathbf{S}\rangle^T) + \alpha\mathbf{A} + \beta . \tag{12}$$

We can write down an iterative update rule for $A_{dm} > 0$ using the Kuhn-Tucker condition $L_{dm}A_{dm} = 0$ [Luenberger 1984] together with the result for the Lagrange multipliers:

$$A_{dm}^{(k+1)} = \frac{[\mathbf{\Sigma}^{-1}\mathbf{X}\langle\mathbf{S}\rangle^T]_{dm}}{[\mathbf{\Sigma}^{-1}\mathbf{A}^{(k)}\langle\mathbf{S}\mathbf{S}^T\rangle]_{dm} + \alpha A_{dm}^{(k)} + \beta} A_{dm}^{(k)} . \tag{13}$$

In the case of no prior knowledge i.e. $\alpha = 0$ and $\beta = 0$, we get a update rule similar to the image space reconstruction algorithm used in positron emission tomography (see e.g. [Pierro 1993] for references) or the more recently proposed non-negative matrix factorization procedure of [Lee and Seung 1999].

3 Mean Field Theory

We will present three different mean field approaches that give us estimates of the source second moment matrix of increasing quality: First, we derive mean field equations using the standard variational mean field theory. Next, using linear response theory, we obtain directly from the variational solution improved estimates of $\langle\mathbf{S}\mathbf{S}^T\rangle$ needed for estimating $\mathbf{A}$ and $\mathbf{\Sigma}$. Finally, we present the adaptive TAP approach of Opper and Winther [Opper and Winther 2000b] which goes beyond the simple factorized trial distribution of variational mean field theory to give a theory which is self-consistent to within linear response corrections. From mean field theory we also get an approximation to the likelihood $P(\mathbf{X}|\mathbf{A}, \mathbf{\Sigma})$ which can be used for model selection [Hansen 2000].[1] In appendix A, we summarize all mean field equations and give an EM-type recipe for solving them.

The following derivation is valid for any source prior without temporal correlations. Specific source priors are discussed in section 4. Although equations for the mean field estimates of the mean and covariance of the sources are written with equality in this section, it is to be understood that they are only approximations.

[1] The variational approximation is a lower bound to the exact likelihood whereas the TAP and LR aproximations — not given here — are not bounds, but hopefully more accurate.

3.1 Variational Approach

We adopt a standard variational mean field theoretic approach and approximate the posterior distribution, $P(\mathbf{S}|\mathbf{X},\mathbf{A},\mathbf{\Sigma})$, in a family of product distributions $Q(\mathbf{S}) = \prod_{mt} Q(S_{mt})$.[2] For a Gaussian likelihood $P(\mathbf{X}|\mathbf{A},\mathbf{\Sigma},\mathbf{S})$, the optimal choice of $Q(S_{mt})$ is given by a Gaussian times the prior [Csató et al. 2000]:

$$Q(S_{mt}) \propto P(S_{mt}) e^{-\frac{1}{2}\lambda_{mt} S_{mt}^2 + \gamma_{mt} S_{mt}} \ . \tag{14}$$

To simplify the notation in the following we will parameterize the likelihood as,

$$P(\mathbf{X}|\mathbf{A},\mathbf{\Sigma},\mathbf{S}) = P(\mathbf{X}|\mathbf{J},\mathbf{h},\mathbf{S}) = \frac{1}{C} e^{-\frac{1}{2}\mathrm{Tr}(\mathbf{S}^T\mathbf{J}\mathbf{S}) + \mathrm{Tr}(\mathbf{h}^T\mathbf{S})} \ , \tag{15}$$

where $\log C = \frac{N}{2}\log\det 2\pi\mathbf{\Sigma} + \frac{1}{2}\mathrm{Tr}\,\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X}$, the $M \times M$ interaction matrix $\mathbf{J}$ and the field $\mathbf{h}$ (having same dimension as $\mathbf{S}$) are given by

$$\mathbf{J} = \mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{A} \tag{16}$$

$$\mathbf{h} = \mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{X} \ . \tag{17}$$

Note that $\mathbf{h}$ acts as an external field from which all moments of the sources can be obtained. This is the key property that we will make use of in the next section when we derive the linear response corrections. The starting point of the variational derivation of mean field equations is the Kullback-Leibler divergence between the product distribution $Q(\mathbf{S})$ and the true source posterior, i.e.

$$\begin{aligned} KL &= \int d\mathbf{S} Q(\mathbf{S}) \log \frac{Q(\mathbf{S})}{P(\mathbf{S}|\mathbf{X},\mathbf{A},\mathbf{\Sigma})} \\ &= \log P(\mathbf{X}|\mathbf{A},\mathbf{\Sigma}) - \log P(\mathbf{X}|\mathbf{A},\mathbf{\Sigma},\mathrm{NMF}) \qquad (18) \\ \log P(\mathbf{X}|\mathbf{A},\mathbf{\Sigma},\mathrm{NMF}) &= \sum_{mt} \log \int dS_{mt} P(S_{mt}) e^{-\frac{1}{2}\lambda_{mt} S_{mt}^2 + \gamma_{mt} S_{mt}} \qquad (19) \\ &\quad + \frac{1}{2}\sum_{mt} (\lambda_{mt} - J_{mm}) \langle S_{mt}^2 \rangle + \mathrm{Tr}(\mathbf{h} - \mathbf{\gamma})^T \langle \mathbf{S} \rangle \\ &\quad + \frac{1}{2}\mathrm{Tr}\langle \mathbf{S}^T \rangle (\mathrm{diag}(\mathbf{J}) - \mathbf{J}) \langle \mathbf{S} \rangle - \ln C \ , \end{aligned}$$

where $P(\mathbf{X}|\mathbf{A},\mathbf{\Sigma},\mathrm{NMF})$ is the naive mean field approximation to the likelihood and $\mathrm{diag}(\mathbf{J})$ is the diagonal matrix of $\mathbf{J}$. The Kullback-Leibler is zero when $P = Q$ and positive otherwise. The parameters of Q should consequently be chosen as to minimze KL. The saddle points define the mean field equations:[3]

$$\frac{\partial KL}{\partial \langle \mathbf{S} \rangle} = 0 : \qquad \mathbf{\gamma} = \mathbf{h} - (\mathbf{J} - \mathrm{diag}(\mathbf{J}))\langle \mathbf{S} \rangle \tag{20}$$

$$\frac{\partial KL}{\partial \langle S_{mt}^2 \rangle} = 0 : \qquad \lambda_{mt} = J_{mm} \ . \tag{21}$$

[2] Note that $Q(S_{mt})$ is a also the variational mean field approximation to the marginal distribution $\int \prod_{m' \neq m, t' \neq t} dS_{m't'} P(\mathbf{S}|\mathbf{X},\mathbf{A},\mathbf{\Sigma})$.

[3] The requirement that we should be at a local minima of $\log P(\mathbf{X}|\mathbf{A},\mathbf{\Sigma},\mathrm{NMF})$ is fulfilled when the covariance matrix eq. (25) is positive definite. To test whether we are at the global minima is harder. However, when the model is well-matched to the data, we expect the problem to be convex.

The remaining two equations depend explicitly on the source prior, $P(\mathbf{S})$;

$$\frac{\partial KL}{\partial \gamma_{mt}} = 0: \qquad \langle S_{mt} \rangle = \frac{\partial}{\partial \gamma_{mt}} \log \int dS_{mt} P(S_{mt}) e^{-\frac{1}{2}\lambda_{mt} S_{mt}^2 + \gamma_{mt} S_{mt}} \equiv f(\gamma_{mt}, \lambda_{mt}) \tag{22}$$

$$\frac{\partial KL}{\partial \lambda_{mt}} = 0: \qquad \langle S_{mt}^2 \rangle = -2 \frac{\partial}{\partial \lambda_{mt}} \log \int dS_{mt} P(S_{mt}) e^{-\frac{1}{2}\lambda_{mt} S_{mt}^2 + \gamma_{mt} S_{mt}} \ . \tag{23}$$

The variational mean $f(\gamma_{mt}, \lambda_{mt}$ plays as crucual role in defining the mean field algorithm since all dependence upon the prior is implicit in f (and in $\frac{\partial f}{\partial \gamma}$ as well for the advanced methods). In section 4, we calculate $f(\gamma_{mt}, \lambda_{mt})$ for some of the prior distributions found in the ICA literature.

3.2 Linear Response Theory

So far we have not discussed how to obtain mean field approximations to the covariances

$$\chi^{tt'}_{mm'} \equiv \langle S_{mt} S_{m't'} \rangle - \langle S_{mt} \rangle \langle S_{m't'} \rangle \ .$$

Since variational mean field theory uses a factorized trial distribution, the covariances between different variables is trivially predicted to be zero. However, using linear response theory, we can improve the variational mean field solution. As mentioned earlier, $\mathbf{h}$ acts as an external field. This makes it possible to calculate the means and covariances as derivatives of $\log P(\mathbf{X}|\mathbf{J}, \mathbf{h})$, i.e.

$$\langle S_{mt} \rangle = \frac{\partial \log P(\mathbf{X}|\mathbf{J}, \mathbf{h})}{\partial h_{mt}} \tag{24}$$

$$\chi^{tt'}_{mm'} = \frac{\partial^2 \log P(\mathbf{X}|\mathbf{J}, \mathbf{h})}{\partial h_{m't'} \partial h_{mt}} = \frac{\partial \langle S_{mt} \rangle}{\partial h_{m't'}} \ . \tag{25}$$

These relations are exact when using the exact likelihood. However, we can also use the NMF likelihood through the mean field equations (20), (21) and (22) to derive an approximate equation for $\chi^{tt'}_{mm'}$

$$\begin{aligned} \chi^{tt'}_{mm'} &= \frac{\partial f(\gamma_{mt}, \lambda_{mt})}{\partial \gamma_{mt}} \frac{\partial \gamma_{mt}}{\partial h_{m't'}} \\ &= \frac{\partial f(\gamma_{mt}, \lambda_{mt})}{\partial \gamma_{mt}} \left(- \sum_{m'', m'' \neq m} J_{mm''} \chi^{tt}_{m''m'} + \delta_{mm'} \right) \delta_{tt'} \ . \end{aligned} \tag{26}$$

As a direct consequence of the lack of temporal correlations in the present setting, the $\boldsymbol{\chi}$-matrix factorizes in time, i.e. $\chi^{tt'}_{mm'} = \delta_{tt'} \chi^{t}_{mm'}$. We can straightforwardly solve for $\chi^{t}_{mm'}$

$$\chi^{t}_{mm'} = \left[(\boldsymbol{\Lambda}_t + \mathbf{J})^{-1} \right]_{mm'} \ , \tag{27}$$

where we have defined the diagonal matrix

$$\boldsymbol{\Lambda}_t = \mathrm{diag}\left(\Lambda_{1t}, \ldots, \Lambda_{Mt}\right), \qquad \Lambda_{mt} \equiv \left(\frac{\partial f(\gamma_{mt}, \lambda_{mt})}{\partial \gamma_{mt}} \right)^{-1} - J_{mm} \ . \tag{28}$$

For comparison, the naive mean field result is $\chi^{t,\mathrm{NMF}}_{mm'} = \delta_{mm'} \frac{\partial \langle S_{mt} \rangle}{\partial h_{mt}}$ which follows directly from eq. (23).

Why is the covariance matrix obtained by linear response more accurate? Here, we give an argument that can be found in Parisi's book on statistical field theory [Parisi 1988]: Let us assume (as always implicit in any mean field theory) that the approximate and exact distribution is close in some sense, i.e. $Q(\mathbf{S}) - P(\mathbf{S}|\mathbf{X}, \mathbf{A}, \mathbf{\Sigma}) = \varepsilon$. Then by direct application of the factorized distribution we have $\langle S_{mt} S_{m't} \rangle_{\text{Exact}} = \langle S_{mt} S_{m't} \rangle_{\text{NMF}} + \mathcal{O}(\varepsilon)$. On the other hand since KL, eq. (18) is non-negative the NMF theory log-likelihood gives a lower bound on the log-likelihood. Consequently, the linear term vanishes in the expansion of the log-likelihood: $\log P(\mathbf{X}|\mathbf{A}, \mathbf{\Sigma}) = \log P(\mathbf{X}|\mathbf{A}, \mathbf{\Sigma}, \text{NMF}) + \mathcal{O}(\varepsilon^2)$. Obtaining moments of the variables through derivatives of the approximate log-likelihood, i.e. by linear response, is therefore more precise than to use the trial distribution directly.

For some specific cases it is possible to demonstrate the improvement directly. Consider the Gaussian prior[4] $P(S_{mt}) \propto \exp(-S_{mt}^2/2)$. In this case the variational mean field, eq. (22) is given by $f(\gamma, \lambda) = \gamma/(1+\lambda)$. Thus, the variational mean field theory predicts $\chi^{t,\text{NMF}}_{mm'} = \delta_{mm'} \frac{\partial \langle S_{mt} \rangle}{\partial h_{mt}} = 1/(1+\lambda_{mt}) = 1/(1+J_{mm})$. However, the linear response estimate eq. (27) gives $\chi^{t,\text{LR}}_{mm'} = \left[(\mathbf{I}+\mathbf{J})^{-1}\right]_{mm'}$ and hence reconstructs the full covariance matrix identical with the exact result obtained by direct integration.

3.3 Adaptive TAP Approach

So far we have derived two different estimates of the covariance matrix from variational mean field theory: $\chi^{t,\text{NMF}}_{mm'} = \delta_{mm'} \frac{\partial \langle S_{mt} \rangle}{\partial h_{mt}}$ and $\chi^{t,\text{LR}}_{mm'} = \left[(\mathbf{\Lambda}_t + \mathbf{J})^{-1}\right]_{mm'}$. Obviously there is no guarentee that the two estimates are identical. Variational mean field theory is thus not self-consistent to within linear response corrections. The adaptive TAP approach [Opper and Winther 2000b] on the other hand goes beyond the factorized trial distribution and requires self-consistency for the covariances estimated by linear response. This is achieved by introducing a set of MT additional mean field (or variational) parameters, the variances λ_{mt} in the marginal distribution eq. (14), such that the diagonal term $\chi^{t,\text{TAP}}_{mm}$ obeys

$$\frac{\partial \langle S_{mt} \rangle}{\partial h_{mt}} = \left[(\mathbf{\Lambda}_t + \mathbf{J})^{-1}\right]_{mm} \tag{29}$$

where Λ_{mt} and γ_{mt} now depend upon λ_{mt}:

$$\Lambda_{mt} = \left(\chi^t_{mm}\right)^{-1} - \lambda_{mt} \tag{30}$$

$$\gamma_{mt} = h_{mt} - \sum_{m'} (J_{mm'} - \lambda_{m't}\delta_{mm'}) \langle S_{m't} \rangle \ . \tag{31}$$

To recover the variational mean field equations (28) and (20), we just let $\lambda_{mt} = J_{mm}$. It is beyond the scope of this paper to rederive the adaptive TAP mean field theory, consult [Opper and Winther 2000b] for a derivation valid for models with quadratic interactions and general variable prior. However, we have chosen to present and test the resulting theory because it offers the most advanced (and hopefully the most precise) mean field approximation for this type of model.

4 Source Models

In this section we calculate for various source priors the variational mean f, eq. 22) and the derivative $\partial f/\partial \gamma$ needed for the linear response correction and adaptive TAP.

[4] It is noted that a Gaussian source prior is not suitable for doing source separation. We merely use it here to shown that the linear response correction in this case recovers the exact result.

The priors that we are considering are all chosen such that the variational mean can be calculated using tables of standard integrals, e.g. [Gradshteyn and Ryzhik 1980]. It turns out to be convenient to introduce the Gaussian kernel D with unit variance and its associated cumulative distribution function (cdf.) Φ in order to keep the following expressions of a manageable size, i.e.

$$D(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) , \qquad D'(x) = -xD(x) \tag{32}$$

$$\Phi(x) = \int_{-\infty}^{x} D(t)dt , \qquad \Phi'(x) = D(x) . \tag{33}$$

4.1 Summary of source priors

Table 1 summarizes the variational means and response functions corresponding to the priors described in this paper. It should be mentioned that this is by no means a complete list of all priors for which it is possible to calculate these quantities, e.g. the Rayleigh distribution is one such prior.

Source Prior	$P(S)$	Mean Function $f(\gamma,\lambda) = \langle S \rangle$	Response Func. $\frac{\partial \langle S \rangle}{\partial \gamma}$
Binary	$\frac{1}{2}\delta(S-1) + \frac{1}{2}\delta(S+1)$	$\tanh(\gamma)$	$1 - \langle S \rangle^2$
Gaussian Mix.	eq. (34)	eqs. (36) & (38)	
Gaussian	$\frac{1}{\sqrt{2\pi}} \exp(-S^2/2)$	$\gamma/(1+\lambda)$	$1/(1+\lambda)$
Heavy tail	not analytic	$\frac{\gamma}{\lambda} - \alpha\frac{\gamma}{\lambda\alpha+\gamma^2}$	$\frac{1}{\lambda} + \alpha\frac{\gamma^2-\lambda\alpha}{(\lambda\alpha+\gamma^2)^2}$
Uniform	$\frac{1}{b-a}\Theta(S-a)\Theta(b-S)$	eq. (62)	eq. (63)
Laplace	$\frac{1}{2}\exp(-\lvert S\rvert)$	eq. (40)	eq. (42)
Pos. Gauss	$\sqrt{\frac{2}{\pi}}\exp(-S^2/2)\Theta(S)$	eq. (57)	eq. (59)
Exponential	$\exp(-S)\Theta(S)$	eq. (44)	eq. (45)

Table 1: The variational mean and response function corresponding to various source priors. The three first rows describe source priors having negative, zero and positive kurtosis, respectively. The fourth row express non-negative priors. The step–function is defined as $\Theta(S) = 1$ for $S > 0$ and zero otherwise.

4.2 Mixture of Gaussians source prior

In this section we consider a general mixture of Gaussians, i.e.

$$p(S|\mu,\sigma) = \sum_{i=1}^{N_m} \pi_i p(S|\mu_i,\sigma_i) , \qquad S \in \mathbb{R} \tag{34}$$

where each of the N_m individual mixture components are parametrized by,

$$p(S|\mu_i,\sigma_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{1}{2}(S-\mu_i)^2/\sigma_i^2} . \tag{35}$$

Using this source prior the generative ICA model becomes the independent factor analysis model proposed in [Attias 1999]. Since the main scope of this paper is concerned

with reliable inferring mean sufficient statics wrt. the sources we will in contrary to [Attias 1999] always regard the source parameters as fixed, e.g. we are at no times adapting the source priors to data. However, it is straight forward to extend the proposed methodology to allow for this possibility, e.g. in a EM setting where the improved mean field solutions are being used in the posterior expectation of the complete log likelihood.

Trivial but tedious calculations shows that the variational mean f of a mixture of Gaussians is given by,

$$f = \frac{\sum_{i=1}^{N_m} \kappa_i \frac{\gamma\sigma_i^2+\mu_i}{\lambda\sigma_i^2+1} e^{\xi_i}}{\sum_{i=1}^{N_m} \kappa_i e^{\xi_i}} \ , \tag{36}$$

where we have introduced

$$\kappa_i = \frac{\pi_i}{\sqrt{\lambda\sigma_i^2+1}} \ , \quad \text{and} \quad \xi_i = -\frac{1}{2}((\mu_i/\sigma_i)^2 - \frac{(\gamma\sigma_i+\mu_i/\sigma_i)^2}{\lambda\sigma_i^2+1}) \ . \tag{37}$$

The derivatives wrt. γ are easy to obtain but are left out in the interest of space. For the special case of a mixture of two Gaussians ($N_m = 2$) with common variance σ^2 and means $\mu_i = \pm\mu$ we get,

$$f = \frac{1}{\lambda\sigma^2+1}(\gamma\sigma^2 + \mu \tanh(\frac{\gamma\mu}{\lambda\sigma^2+1})) \ . \tag{38}$$

For $\sigma^2 = 0$ and $\mu = 1$, we recover the variational mean for the binary source $P(S) = \frac{1}{2}\delta(S-1) + \frac{1}{2}\delta(S+1)$: $f = \tanh(\gamma)$. This particular choice of the bi-Gaussian source distribution (eq. 38) which is also known as the symmetric Pearson mixture density, was proposed in [Girolami 1998] as a simple way of archiving a negative kurtosis (sub-Gaussian) density function. To become familiar with the f-function and its derivative, consider the variational mean of the bi-Gaussian with $\sigma^2 = 1$ shown in figure 1(a,b) for two values of μ; namely $\mu = 1$, for which the density function is uni-modal and $\mu = 4$ for which the density function is significantly bimodal. We seen that the more bimodal the source distribution are the more compact becomes the region of high curvature. By introducing additional mixture components it is possible to form the region of high curvature, which is illustrated in figure 1(g) in the case of a mixture of five Gaussians.

4.3 Laplace source prior

Although a sub-Gaussian distribution may be a reasonable source prior for some applications, e.g. telecommunications (discrete priors, see e.g. [van der Veen 1997]) or processing of functional magnetic resonance images [Petersen et al. 2000], there is, however, a large class of interesting real world signals, such as speech, which have heavier tails than the Gaussian distribution. We therefore need to consider source priors which have positive kurtosis (super-Gaussian). One such choice which have been widely used in the ICA community is $P(S) = 1/(\pi \cosh S)$ [Bell and Sejnowski 1995; MacKay 1996]. Using this prior, however, it is not possible to calculate the variational mean analytically. Instead we consider the Laplace or double exponential distribution which is very similar. The Laplace density is given by,

$$p(S) = \frac{\eta}{2} e^{-\eta|S|} \ , \qquad S \in \mathbb{R}, \ \eta > 0. \tag{39}$$

The variational mean can be calculated as,

$$f = \frac{1}{\sqrt{\lambda}} \frac{\xi_+\kappa_+ + \xi_-\kappa_-}{\kappa_+ + \kappa_-} \tag{40}$$

where we have introduced,

$$\xi_\pm = \frac{\gamma \mp \eta}{\sqrt{\lambda}} , \quad \text{and} \quad \kappa_\pm = \Phi(\pm\xi_\pm) D(\xi_\mp) . \tag{41}$$

Using eqs. (32) and (33), the derivative is found to be,

$$\frac{\partial f}{\partial \gamma} = \frac{1}{\lambda} \left(1 - \xi_- \xi_+ + D(\xi_+) D(\xi_-) \frac{\xi_+ - \xi_-}{\kappa_+ + \kappa_-} + \sqrt{\lambda} \frac{(\xi_+ \kappa_- + \xi_- \kappa_+)}{(\kappa_+ + \kappa_-)} f \right) . \tag{42}$$

Figure 1(c,d) shows the variational mean and its derivative for a slowly decaying ($\eta = 0.5$) and a fast decaying ($\eta = 2$) Laplacian prior. The Laplacian prior have, contrary to the bi-Gaussian source, its region of high curvature for numerical large values of γ.

4.4 Exponential source prior

Some application domains naturally restrict the possible range of the hidden sources and the mixing matrix due to the physical interpretation of these quantities in the generative model. This is for instance the case when the measured signal is known to be a positive superposition of latent counting numbers or intensities. Positivity constrains are relevant, e.g., in "parts based representations" of natural images, deconvolution of the power spectrum of nuclear magnetic resonance resonance (NMR) spectrometers and latent semantic analysis in text mining [Lee and Seung 1999]. In this section we consider the exponential source prior parameterized by,

$$p(S) = \eta e^{-\eta S} , \qquad S \in \mathbb{R}_+ , \ \eta > 0 \tag{43}$$

which gives

$$f = \frac{1}{\sqrt{\lambda}} \frac{\xi \Phi(\xi) + D(\xi)}{\Phi(\xi)} \tag{44}$$

$$\frac{\partial f}{\partial \gamma} = \frac{1}{\lambda} + \frac{D(\xi)}{\sqrt{\lambda} \Phi(\xi)} f . \tag{45}$$

with

$$\xi = \frac{\gamma - \eta}{\sqrt{\lambda}} . \tag{46}$$

Figure 1(e,f) shows the variational mean and its derivative for the exponential source prior. It is verified that the exponential variational mean is non-negative. At this point we will make some short remarks on some implementational issues when the normal cdf. Φ appears in the denominator of the variational mean. Special care have to be taken when $\xi \to -\infty$, e.g. when $\gamma - \eta < 0$ and λ is small, i.e. for small self-interactions. Using l'Hospital's rule together with eqs. (32) and (33), it is seen that

$$\frac{D(\xi)}{\Phi(\xi)} \to -\xi \quad \text{for} \quad \xi \to -\infty , \tag{47}$$

which in turn implies that the variational mean $f \to 0$ and its derivative $(\partial f / \partial \gamma) \to 1/\lambda$ for $\xi \to -\infty$. In section 5.4, we will use this prior to learn a set of sparse localized basis functions in images. The source priors considered until now are just some examples of priors where the variational mean can be computed analytically. In appendix B we simply state some additional examples of priors for which this calculation can be carried out analytically.

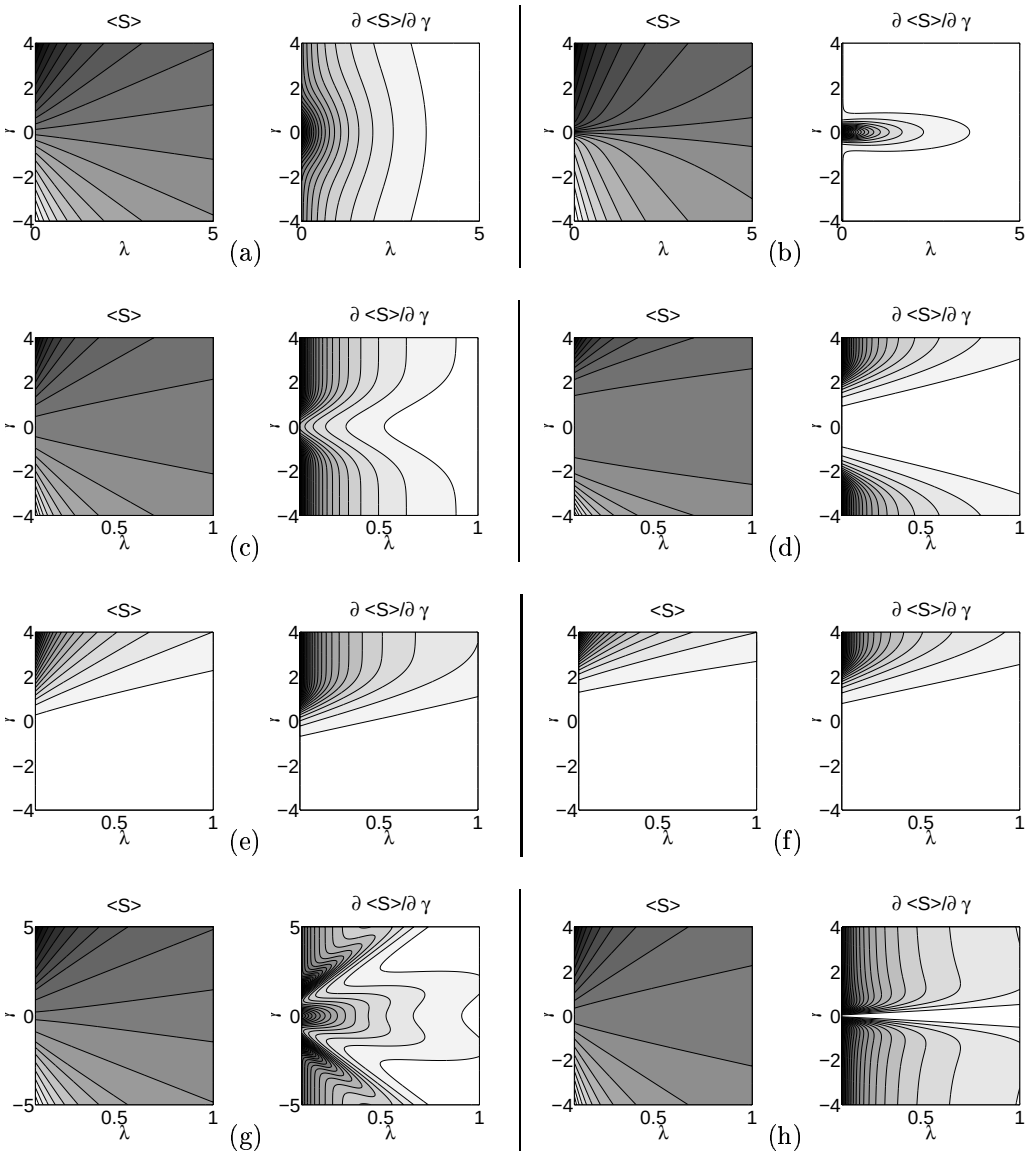


Figure 1: Shows the variational mean f (left row) and its derivative f' (right row) as a function of γ and λ. (a) and (b) shows the bi-Gaussian case with $\sigma^2 = 1$ for $\mu_i = \pm 1$ and $\mu_i = \pm 4$, respectively. (c) and (d) shows the Laplacian prior for decay rates $\eta = 1/2$ and $\eta = 2$, respectively. (e) and (f) shows the exponential prior for decay rates $\eta = 1/2$ and $\eta = 2$, respectively.; (g) shows the variational mean f and the derivative f' of a mixture of five Gaussian with mixing proportions $\pi_i = 1/5$, means $\mu_i = \{-4, -1, 0, 1, 4\}$ and standard deviations $\sigma_i = \{1, 2, 4, 2, 1\}$. (h) shows the heavy tailed prior eq. (48) with $\alpha = 1$.

4.5 Power law tail prior

In the previous sections we have only considered source priors for which it was possible to carry out the integration eq. (22) analytically. For arbitrary source priors, however, the one dimensional integral may be be solved using standard approaches for numerical integration. Alternatively, we could simply use the insight gained in the previous sections, where we considered the functional form of the variational mean of various source priors, to come up with computationally tractable f functions directly. To give an example of this, we will construct a f which for large $|\gamma|/\sqrt{\lambda}$ corresponds to a distribution with a power law tail $P(S) \propto |S|^{-\alpha}$ for $|S|$ large. In this limit the integral in eq. (22) is dominated by its saddlepoint. The saddlepoint value of S is $S_0 = \frac{\gamma}{2\lambda}(1+\sqrt{1-\frac{4\alpha\lambda}{\gamma^2}}) \approx \frac{\gamma}{\lambda} - \frac{\alpha}{\gamma}$. This gives the behavior of the mean function for large γ. We can now straightforwardly construct a mean function that has this assumptotic behavior and is weel-defined for small values of γ:

$$f = \frac{\gamma}{\lambda} - \frac{\alpha\gamma}{\alpha\lambda + \gamma^2} \; . \tag{48}$$

Figure 1(h) shows the heavy tail f-function as a function of γ and λ. Figure 2 shows for a fixed $\lambda = 1$ the variational mean and derivative for some of the unconstrained source priors considered so far. For $\gamma \to \infty$, the Gaussian and the uniform (improper) prior give respectively the the lower and upper value for f for the priors considered.

The variational means and derivatives for the priors considered in this paper are summarized in the table in section 4.1.

5 Simulations

In this section we compare the performance of the different mean field approaches described in the previous sections, i.e. NMF, LR correction and TAP. To begin with, we conduct two experiments with artificial generated data. The source priors used in these experiments are equal to the source prior which generated the dataset. We consider both the complete case in which 2 binary sources are mixed into 2 sensors and the overcomplete case of 3 continuous sources mixed into 2 sensors. Finally, we apply the linear response corrected mean field approach for to perform ICA on two real world datasets; namely speech signals and parts of the MNIST handwritten digit database.

5.1 Synthetic binary sources in an complete setting

Independent component analysis of binary sources have been considered e.g. in data transmission using binary modulation schemes such as MSK or biphase codes [van der Veen 1997]. Here, we consider a binary source $S = \{\pm 1\}$ with prior distribution $\frac{1}{2}[\delta(S-1)+\delta(S+1)]$. In this case we recover the well known mean field equations $\langle S \rangle = \tanh(\gamma)$. Figure 3(a) shows the column vectors of the mixing matrix and 1000 samples generated from the ICA generative model using a fairly low noise variance, $\sigma^2 = 0.3$. Ideally, the noise-less measurements would consist of the four combinations (with sign) of the columns in the mixing matrix. However, due to the noise, the measurements will be scattered around these prototype observations (shown as $+$ in figure 3(a)). Figure 3(b) shows, for each of the mean field approaches, the variance as a function of iteration number. At these moderate noise variances an improvement in the convergence rate is obtained by using the linear response corrected mean field solution. The adaptive TAP approach, on the other hand, is seen to have a slower convergence rate and only a marginal improvement in the estimated noise variance and mixing matrix is obtained.

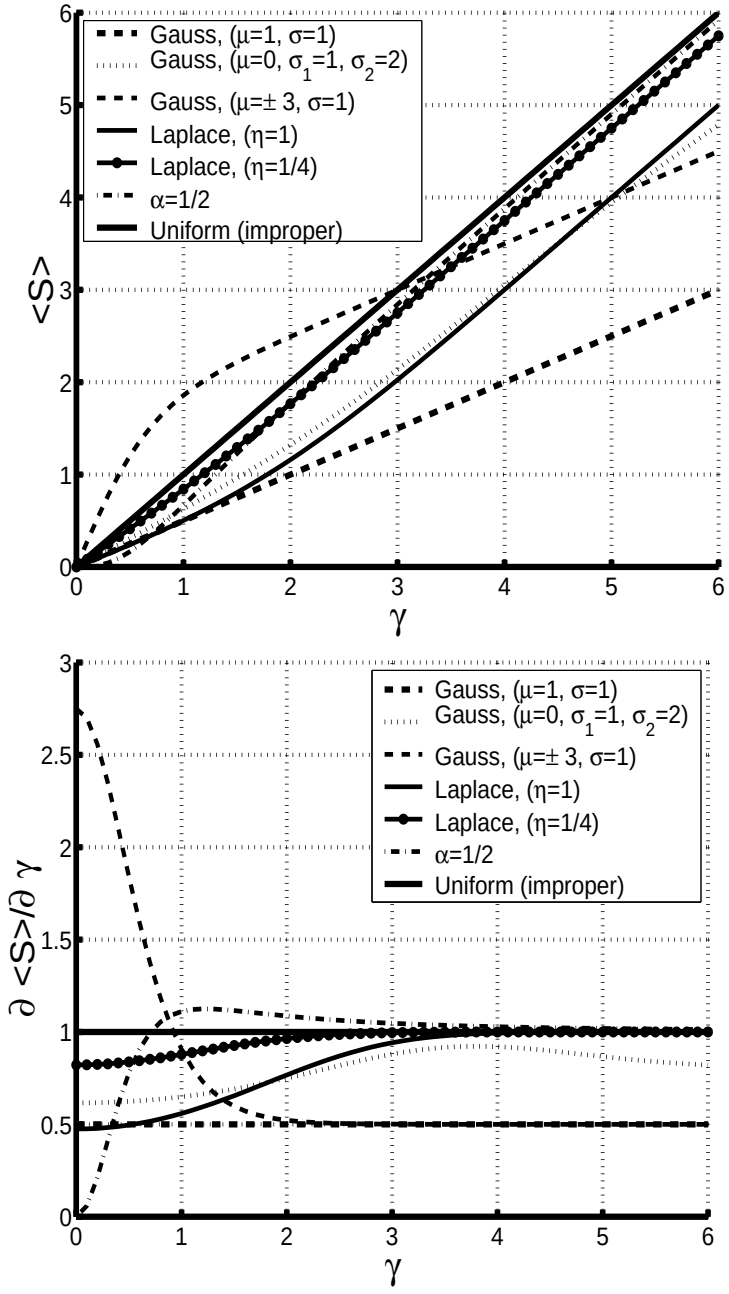


Figure 2: Shows the variational mean (top) and derivative (lower) as a function of γ for various source priors and fixed $\lambda = 1$. From top to bottom the legends are; [- -] Gaussian with unit mean and variance; [$\cdots$] Mixture of two Gaussians with 0 mean and std. 1 and 2; [- -] Mixture of two Gaussians with unit variance and mean at ± 3; [—] and [-•-] Laplacian with $\eta = 1$ and $\eta = 1/4$, respectively; [·-·-] Heavy tail with $\alpha = 1/2$; [- -] Uniform (improper) distribution.

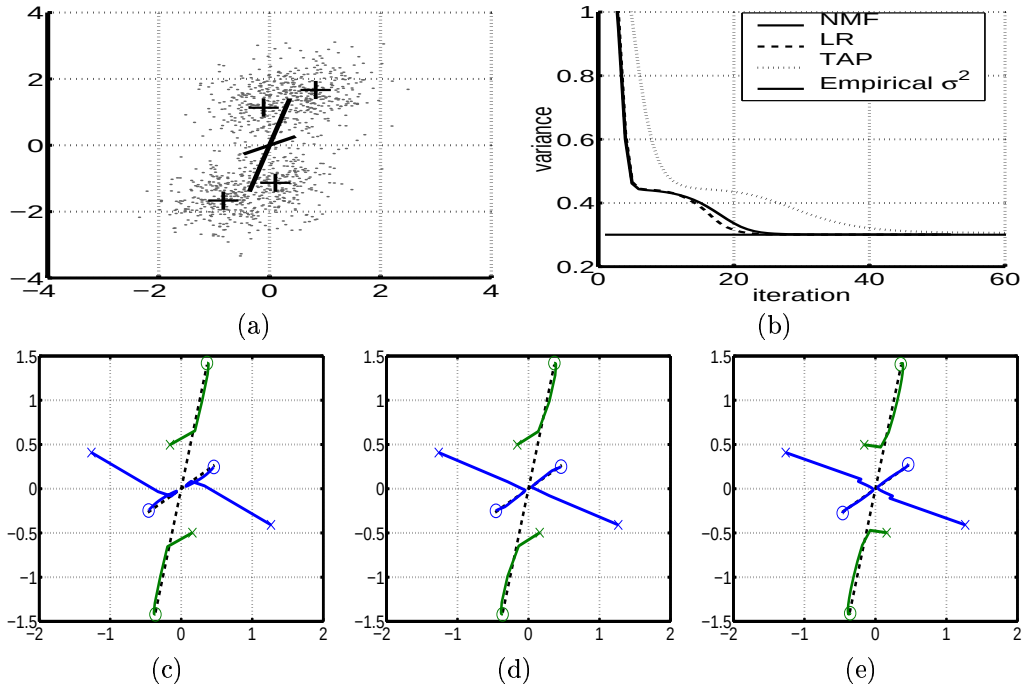


Figure 3: Binary source recovery for a low noise variance, $\sigma^2 = 0.3$. (a) Shows 1000 measurements (scatter plot), +/- the column vectors of the true mixing matrix (the solid axis) and the measurement prototypes (+) for the noise-less case. (b) Shows the estimated variance for NMF, LR and TAP as a function of iteration. The thick solid line is the true empirical noise variance. The empirical variance is the variance of the 1000 random noise contributions. The trajectories of the fix-point iteration using (c) NMF, (d) LR and (e) adaptive TAP. The initial condition is mark '$\times$' and final point '$\circ$'. The dashed lines are the true mixing matrix.

This is due to the fact that this approach is critically sensitive to how well the variational parameters have been determined.

Figure 3(c,d,e) shows, for the different mean field approaches, the trajectories of the fix-point iterations. All the methods uses the same initial conditions ('$\times$') and the final point in the trajectory is mark '$\circ$'. The dashed lines are +/- the column vectors of the true mixing matrix. In this case there is no significant difference in the mixing matrix estimated using the different mean field approaches.

We now increase the noise variance to $\sigma^2 = 1$. In this case it is hard to identify the prototype signals from the measured data (see figure 4(a)). The naive mean field approach fails in recovering the mixing matrix. Figure 4(c) shows that one of the directions in the mixing matrix vanishes during the fix-point iterations which in turn results in the noise variance being overestimated (see figure 4(b)). However, the linear response corrected mean field approach and adaptive TAP recovers the true mixing matrix.

5.2 Continuous sources in an overcomplete setting

In this section the problem is to recover more sources than sensors; in particular we consider mixing 3 source into 2 sensors. The source used in this experiment is the

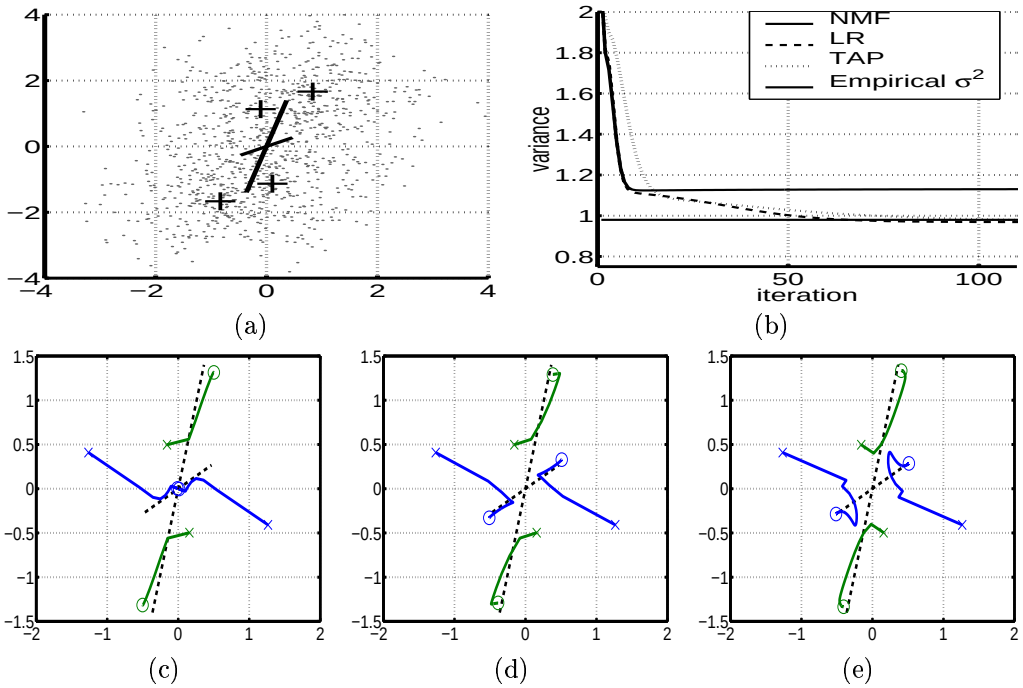


Figure 4: Binary source recovery for a high noise variance, $\sigma^2 = 1$. (a) Shows 1000 measurements (scatter plot), +/- the column vectors of the true mixing matrix (the solid axis) and the measurement prototypes (+) for the noise-less case. (b) Shows the estimated variance for NMF, LR and TAP as a function of iteration. The thick solid line is the true empirical noise variance. The trajectories of the fix-point iteration using (c) NMF, (d) LR and (e) adaptive TAP. The initial condition is mark '×' and final point 'o'. The dashed lines are the true mixing matrix.

symmetric Pearson mixture eq. (38) with $\mu = 1$. A total of 2000 samples was generated from the generative model (see figure 5(a)) and the three mean field approaches was used to learn the mixing matrix. The trajectories plot in figure 5(c) shows that the naive mean field approach fails in recovering the mixing matrix. Similar to the binary case with high variance, one of the directions in the mixing matrix vanishes (see figure 5). Only the dominant direction in the dataspace is captured whereas the two remaining direction collapses into one "mean" direction. However, both the linear response corrected and the adaptive TAP mean field approaches succeed in estimating the mixing matrix. We will restrict ourselves to the LR approach in the next real world examples since NMF has turned out to fail in some cases and TAP is considerably more computationally expensive while giving comparable performance.

5.3 Separating 3 speakers from 2 microphones

In this section we consider the problem of separating three speakers from two microphones. At hand we have the three original speech signals, each having a duration of 1 second and sampled at 8 kHz. The speech signals is then instantaneously linearly mixed into 2 microphones. Figure 6(a) shows a scatter plot of the 8000 samples in the measurement (microphone) space. The fact that natural speech has a heavy tailed distri-

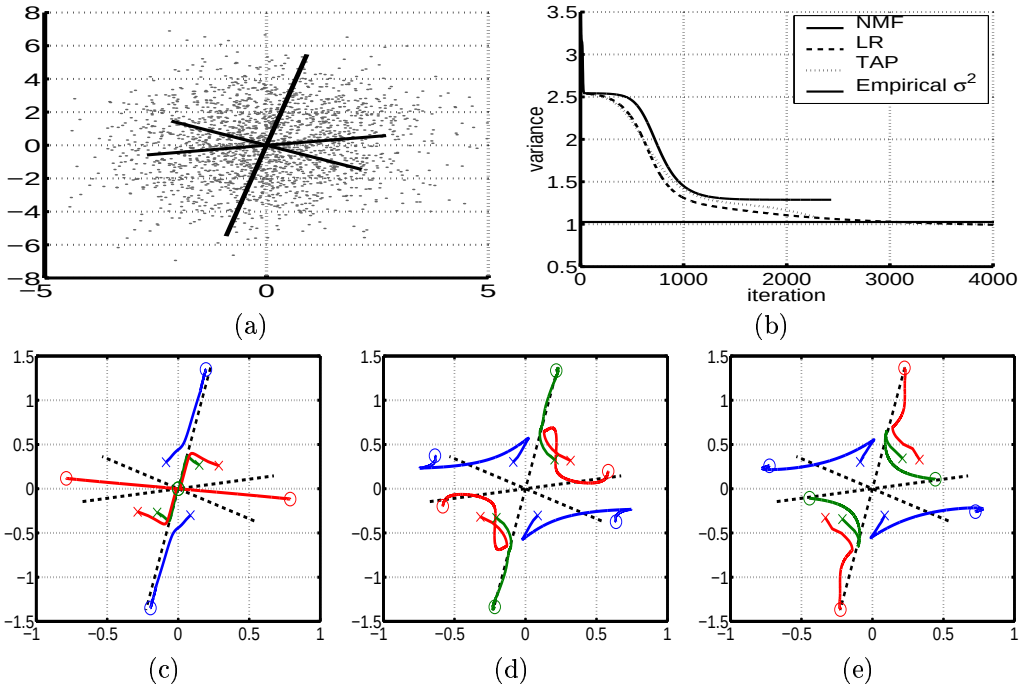


Figure 5: Overcomplete continuous source recovery with $\sigma^2 = 1$. (a) Shows 2000 measurements (scatter plot), +/- the column vectors (4 times axis) of the true mixing matrix (the solid axis). (b) Shows the estimated variance for NMF, LR and TAP as a function of iteration. The thick solid line is the true empirical noise variance. The trajectories of the fix-point iteration using (c) NMF, (d) LR and (e) adaptive TAP. The initial condition is mark '×' and final point 'o'. The dashed lines are the true mixing matrix.

bution makes this overcomplete problem somewhat easiler in the sense that the hidden directions of the mixing matrix reviels themself clearly in the scatter plot. The linear response corrected mean field approach was used in performing ICA with the computationally tractable variational mean eq. (48) with $\alpha = 1$. The initial mixing matrix was randomly picked (shown as the dotted axis in figure 6(a)). Figure 6(b) shows the convergence of the algorithm in term of the angle between the estimated directions and the true directions (the dashed lines in figure 6(a)). Figure 6(a) shows that the algorithm converges rapidly to a mixing matrix which is close to the one that gave raise to the mixed speech signals. each of the inferred sources against each of the true sources (see three recoverd sources is nicely correlated with exactly one of the true sources and (more or less) uncorrelated with the remaning sources (note that the solution is invariant under a relabelling of the sources and columns of the mixing matrix plus a change of scale and sign).

5.4 Local feature extraction with sparse positive encoding

In this section we apply the linear response corrected ICA algorithm to the problem of finding a small set of localized images representing parts of the digit images in the MNIST handwritten digit database. For illustration purposes we will only consider a small subset of the database, namly the first 500 cases of the handwritten digit "3". As mention

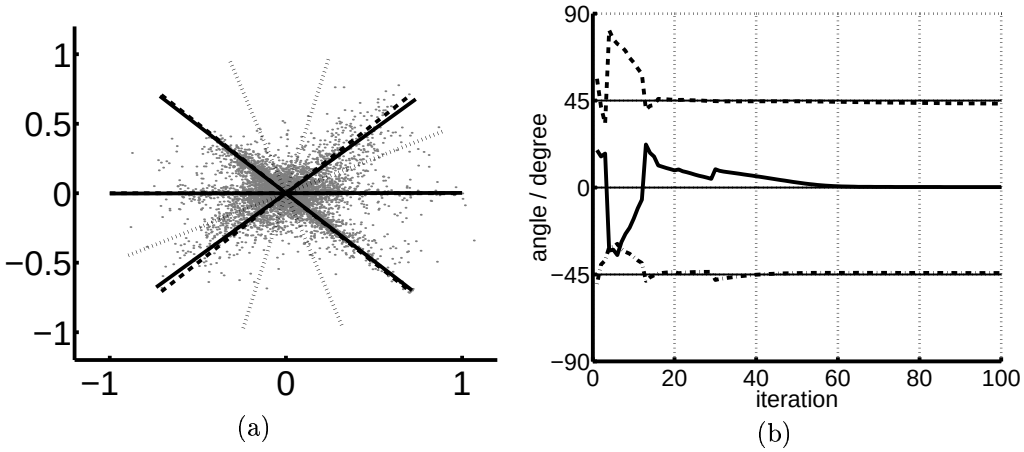


Figure 6: Overcomplete speech separation (3-in-2) using the heavy tailed f-function eq. (48) with $\alpha = 1$; see figure 1(h). (a) scatter plot of 1 sec. of the mixed speech (@8 kHz), the true A (dashed lines), the initial A (black dotted) and the estimated A. (b) shows the estimated angle as a function iteration. The horizontal lines illustrate true angles at 0 and ± 45 degrees.

already in section 4.4 it is natural to consider positive constraints on latent variables (say pixels) when dealing with images. However, such constraints are usually ignored by most of the commonly used preprocessing models e.g. the principal component analysis (PCA) generative model which simply amounts to sequentially finding orthogonal directions (components) with maximum variance in the data space. Ignoring such constraints is problematic since for an unconstrained model to yield positive digit images there have to an interaction between positive and negative regions in different components and it is therefore not obvious what the set of components represents visually.

To illustrate these points we conduct two ICA experiments using the exponential prior $P(S) = e^{-S}$, $S \in \mathbb{R}_+$. In the first experiment we do not constrain the mixing matrix thereas in the second experiment the mixing matrix is constrained to be positive. For both experiment we assume that there are 25 hidden images. Figure 8(a) shows the 25 hidden images obtained using ICA with positively constrained sources but unconstrained mixing matrix. Although the sources in this case are positively constrained, the fact that hidden images are allowed to be subtracted in order to obtain a positive image leads to non-local hidden images which are hard to interpret visually. Figure 8(b) shows the 25 hidden images obtained by performing ICA which inforces the positive constraint on the mixing matrix. In this case the hidden images clearly represents local features, in particular the different handwriting styles/strokes in the various parts of the written digit.

6 Conclusion

In this paper, we have presented a probabilistic (Bayesian) approach to ICA. Sources are estimated by their posterior mean while maximum a posteriori estimates are used for the mixing matrix and the noise covariance. By this procedure we dervied an EM-type algorithm. The expectation step is carried out using different mean field (MF) approaches namely variational (aka ensemble learning or naive MF), linear response and

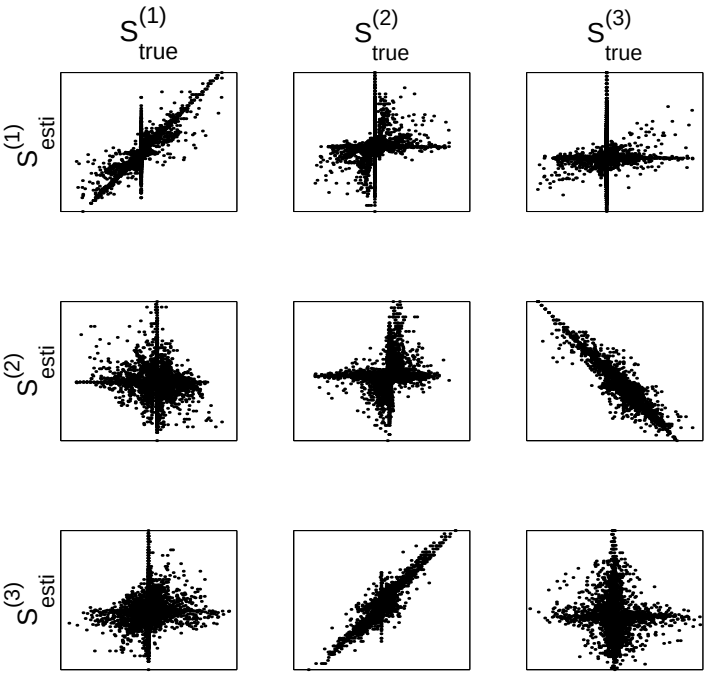


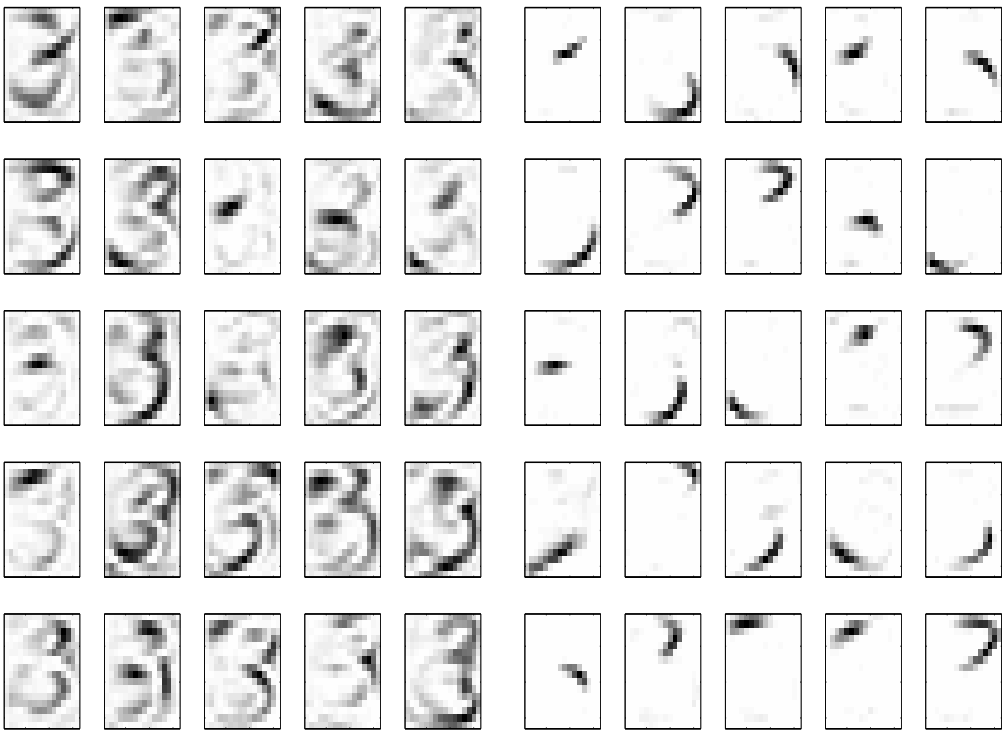
Figure 7: Overcomplete speech separation (3-in-2) using the heavy tailed f-function eq. (48) with $\alpha = 1$. Shows the scatterplots of the ICA estimated sources $S^{(i)}_{esti}$ versus the true sources $S^{(i)}_{true}$, $i = 1, 2, 3$.

adaptive TAP. The MF theories produce estimates of posterior source correlations of increasing quality. These are needed for the maximization step in the estimate for the mixing matrix and the noise.

The importance of a good estimate of correlations is seen for for specific examples where in fact the simplest variational approach fails. The general applicability of the formalism and its MF implementation is demonstrated on local feature extraction in images (using non-negative mixing matrix and source priors) and in overcomplete separation of speech (using heavy tailed source priors). The good performance of the mean field approach supports the belief that we get fair estimates of the posterior means and covariances. However, a rigourous test requires either explicit numerical integration which is possible only for low dimensional problems or Monte Carlo sampling (which may also be inaccurate in complex cases).

In the following, we will discuss a number of possible extensions of this work. One obvious extension is the modelling of temporal correlations. The most general formulation of the model with temporal correlation leads to the consideration of the junction tree algorithm. We are currently working on a mean field algorithm for online belief propagation on the junction tree [Højen-Sørensen et al. 2001].

Optimization of the hyperparameters of the prior can be performed by extending the current EM algorithm. The mean field approach can also be used to derive leave-one-out estimators [Opper and Winther 2000a; Opper and Winther 2000b] that can be used both for optimization of hyperparameters and model selection. Model selection can also be performed using the (approximate mean field) likelihood of a test set.



(a) (b)

Figure 8: Feature extraction of the handwritten digit "3" using a exponential prior with $\eta = 1$ and (a) unconstrained mixing matrix and (b) positive constrained mixing matrix.

Finally, it could be interesting to relax some of the basic requirement of model. Firstly, that of statistical independence of the sources. Our formalism can be extended to treat a priori Gaussian correlations between (the non-Gaussian) sources. We should be able estimate these correlations effectively by for example the linear response technique. Secondly, the model can be extended to nonlinear mixing by e.g. introducing a sigmoidal squashing of the mixed signal. This situation can also with some increase in the computational complexity be included in the mean field framework [Opper and Winther 2000b].

Acknowledgments

We thank Michael Jordan and Manfred Opper for helpful discussions. This research is supported by the Danish Research Councils through the THOR Center for Neuroinformatics and by the Center for Biological Sequence Analysis.

A Algorithmic recipe

In table 2, we give an EM recipe for solving the mean field equations and the equations for the mixing matrix and the noise covariance. It is indicated in the table which equations that have been used. Here, we have giving the equations for adaptive TAP. Linear

response theory is obtained by omitting the updating step for λ_{mt}, i.e. by setting $N_\lambda := 0$. Furthermore setting $\chi^t_{mm'} := \delta_{mm'} f'(\gamma_{mt}, \lambda_{mt})$ instead of $\boldsymbol{\chi}^t := (\boldsymbol{\Lambda}_t + \mathbf{J})^{-1}$ leads to the naive mean field algorithm.

In the table, we have given the update rule for the non-negative mixing matrix eq. (13). To get to the unconstrained mixing matrix, the unconstrained update rule eq. (10) should be used.

Note that we use a greedy update step for all variables but the means $\langle \mathbf{S} \rangle$. Especially adaptive TAP is quite sensitive to the choice of the learning rate η. It is therefore made adaptive such that it is increased with a factor of 1.1 if the sum of the squared deviations $\sum_{mt} |\delta\langle S_{mt}\rangle|^2$ decreases compared to the previos update. Otherwise it is decreased with a factor 2. Our experience with the TAP equations also indicates that running with variable number of updates of $\langle \mathbf{S} \rangle$ could be helpful. However, in the simulations described here we kept the number of iterations fixed.

B Some additional analytical source priors

In this appendix we derive the variational mean and response function for some additional analytical source priors which have not been directly used in this paper. We show these calculations in some details since they are of the same type as the one we carried out in deriving the variational mean of the sources in section 4.

B.1 Positively constrained Gaussian source prior

Calculating the variational mean eq. (22) in general involves the calulation of an intergral of the form,

$$\int dS P(S) e^{-\frac{1}{2}\lambda S^2 + \gamma S} \ , \tag{49}$$

where $P(S)$ is the source prior. The source priors considered in this paper are all of such a form that this integral can reparameterized into a integral over a Gauassian kernel. For this reason in is usefull to have at hand an expression for the integral of a Gaussian kernel, i.e,

$$\begin{aligned}
\int_{-\infty}^{x} dS e^{-\frac{1}{2}\lambda S^2 + \gamma S} &= (\sqrt{2\pi} D(\frac{\gamma}{\sqrt{\lambda}}))^{-1} \int_{-\infty}^{x} dS e^{-\frac{1}{2}\lambda (S - \frac{\gamma}{\lambda})^2} && (50)\\
&= (\sqrt{\lambda}\sqrt{2\pi} D(\frac{\gamma}{\sqrt{\lambda}}))^{-1} \int_{-\infty}^{\xi} dS e^{-\frac{1}{2}\lambda S^2} && (51)\\
&= \frac{\Phi(\xi)}{\sqrt{\lambda} D(\frac{\gamma}{\sqrt{\lambda}})} && (52)
\end{aligned}$$

where $\xi = \sqrt{\lambda}(x - \gamma/\lambda)$. The first equality follows from completing squares and introducing the Gaussian pdf., eq. (32). The second equality follows by changing the integration variable whereas the final equality follows by introducing the Gaussian cdf., eq. (33). We can now calculate the following integral,

$$\int_0^{+\infty} dS e^{-\frac{1}{2}\lambda S^2 + \gamma S} = \int_{-\infty}^{+\infty} (\cdot) - \int_{-\infty}^{0} (\cdot) = \frac{1 - \Phi(-\frac{\gamma}{\sqrt{\lambda}})}{\sqrt{\lambda} D(\frac{\gamma}{\sqrt{\lambda}})} \propto \frac{\Phi(\frac{\gamma}{\sqrt{\lambda}})}{\sqrt{\lambda} D(\frac{\gamma}{\sqrt{\lambda}})} \tag{53}$$

where the factor of proportionality is independent of γ. It is remembered that the actual factor of proportionality is not needed in calculating the variational mean,

$$f(\gamma,\lambda) = (\frac{\Phi}{D})^{-1}\frac{\Phi'_\gamma D - \Phi D'_\gamma}{D^2} = (\frac{\Phi}{D})^{-1}\frac{D^2/\sqrt{\lambda}+\gamma/\lambda\Phi D}{D^2} \tag{54}$$

$$= \frac{\gamma}{\lambda} + \frac{D(\frac{\gamma}{\sqrt{\lambda}})}{\sqrt{\lambda}\Phi(\frac{\gamma}{\sqrt{\lambda}})} \; . \tag{55}$$

We can now return to the problem of calculating the variational mean of a positively constrained Gaussian parameterized by,

$$p(S|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}(S-\mu)^2/\sigma^2} \; , \qquad S \in \mathbb{R} \; , \tag{56}$$

where μ and σ^2 are the mean and variance, respectively. Multiplying the source prior onto the Gaussian kernel and identifying terms it is seen that the product can be written as a Gaussian with $\lambda := \lambda + 1/\sigma^2$ and $\gamma := \gamma + \mu/\sigma^2$. Substituting back into eq. (55) we directly optain the variational mean,

$$f(\gamma,\lambda) = \frac{\gamma+\mu/\sigma^2}{\lambda+1/\sigma^2} + \frac{1}{\sqrt{\lambda+1/\sigma^2}}\frac{D(\kappa)}{\Phi(\kappa)} \tag{57}$$

where we have introduced

$$\kappa = \frac{\gamma+\mu/\sigma^2}{\sqrt{\lambda+1/\sigma^2}} \; , \tag{58}$$

and the response function can be readily derived,

$$\frac{\partial f}{\partial \gamma} = \frac{\mu/\sigma^2}{\lambda+1/\sigma^2}\left(1 - \kappa\frac{D(\kappa)}{\Phi(\kappa)} - \left(\frac{D(\kappa)}{\Phi(\kappa)}\right)^2\right) \; . \tag{59}$$

B.2 Uniform source prior

In this section we consider the uniform prior,

$$P(S) = \frac{1}{b-a} \; , \qquad S \in [a;b] \; , \tag{60}$$

where $b \geq a$. By reusing the calculations made in appendix B.1 we directy optain,

$$\int_a^b dS e^{-\frac{1}{2}\lambda S^2+\gamma S} = \int_{-\infty}^b (\cdot) - \int_{-\infty}^a (\cdot) = \frac{\Phi(\kappa_b)-\Phi(\kappa_a)}{\sqrt{\lambda}D(\frac{\gamma}{\sqrt{\lambda}})} \; , \tag{61}$$

where $\kappa_x = \sqrt{\lambda}(x-\gamma/\lambda) = \sqrt{\lambda}x - \frac{\gamma}{\sqrt{\lambda}}$. Here, we have again left out the normalizing constant since it is of no importance in the caluclation of the variational mean,

$$f(\gamma,\lambda) = \frac{\gamma}{\lambda} + \frac{1}{\sqrt{\lambda}}\frac{D(\kappa_a)-D(\kappa_b)}{\Phi(\kappa_b)-\Phi(\kappa_a)} \; , \tag{62}$$

and the response function,

$$\frac{\partial f}{\partial \gamma} = \frac{1}{\lambda}\left(1 + \frac{\kappa_a D(\kappa_a) - \kappa_b D(\kappa_b)}{\Phi(\kappa_b)-\Phi(\kappa_a)} - \left(\frac{D(\kappa_a)-D(\kappa_b)}{\Phi(\kappa_a)-\Phi(\kappa_b)}\right)^2\right) \; . \tag{63}$$

This appendix showed some illustrative exsamples of the caluculation needed in deriving the variational mean and response functions for the source priors considered in this paper.

References

Attias, H. (1999). Independent factor analysis. *Neural Computation*, 11(4):803–851.

Bell, A. J. and Sejnowski, T. J. (1995). An information maximisation approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159.

Belouchrani, A. and Cardoso, J.-F. (1995). Maximum likelihood source separation by the expectation-maximization technique: deterministic and stochastic implementation. In *In Proc. NOLTA*, pages 49–53.

Csató, L., Fokoué, E., Opper, M., Schottky, B., and Winther, O. (2000). Efficient approaches to Gaussian process classification. In Solla, S., Leen, T., and Müller, K.-R., editors, *Advances in Neural Information Processing Systems*, volume 12, pages 251–257. MIT Press.

Girolami, M. (1998). An alternative perspective on adaptive independent component analysis algorithms. *Neural Computation*.

Girolami, M., editor (2000). *Advances in Independent Components Analysis*. Springer-Verlag, Berlin.

Gradshteyn, I. S. and Ryzhik, I. M. (1980). *Tabel of Integrals, Series, and Products*. Academic Press, New York, corrected and enlarged edition edition.

Hansen, L. K. (2000). Blind separation of noisy image mixtures. In Girolami, M., editor, *Advances in Independent Components Analysis*. Springer-Verlag, Berlin.

Højen-Sørensen, P. A. d. F. R., Winther, O., and Hansen, L. K. (2001). In preparation.

Hyvärinen, A. and Karthikesh, R. (2000). Sparse priors on the mixing matrix in independent component analysis. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000)*, pages 477–452, Helsinki, Finland.

Kappen, H. J. and Rodríguez, F. B. (1998). Efficient learning in Boltzmann machines using linear response theory. *Neural Computation*, 10:1137–1156.

Knuth, K. (1999). A bayesian approach to source separation. In Cardoso, J.-F., Jutten, C., and Loubaton, P., editors, *Proceedings of the First International Workshop on Independent Component Analysis and Signal Separation (ICA'99)*, pages 283–288, Aussios, France.

Lappalainen, H. and Miskin, J. W. (2000). Ensemble learning. In Girolami, M., editor, *Advances in Independent Components Analysis*. Springer-Verlag, Berlin.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401.

Lee, T.-W. (1998). *Independent Component Analysis: Theory and Applications*. Kluwer Academic Publishers, Boston.

Lewicki, M. S. and Sejnowski, T. J. (2000). Learning overcomplete representations. *Neural Computation*, 12(2):337–365.

Luenberger, D. G. (1984). *Linear and Nonlinear Programming*. Addison-Wesley, Reading, MA., second edition.

MacKay, D. J. C. (1996). Maximum likelihood and covariant algorithms for independent component analysis. Technical report, University of Cambridge, Cavendish Laboratory. Draft 3.7.

Opper, M. and Winther, O. (2000a). Gaussian processes for classification: Mean field algorithms. *Neural Computation*, 12(11):2655–2684.

Opper, M. and Winther, O. (2000b). Tractable approximations for probabilistic models: The adaptive tap mean field approach. *Phys. Rev. Lett.* Submitted.

Parisi, G. (1988). *Statistical Field Theory.* Addison-Wesley.

Petersen, K. S., Hansen, L. K., Kolenda, T., Rostrup, E., and Strother, S. (2000). On the independent components in functional neuroimages. In Pajunen, P. and Karhunen, J., editors, *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000)*, pages 615–620, Helsinki, Finland.

Peterson, C. and Anderson, J. R. (1987). A mean field theory learning algorithm for neural networks. *Complex Systems*, 1:995–1019.

Pierro, A. R. (1993). On the relation between the ISRA and the EM algorithm for positron emission tomography. *IEEE Transactions on Medical Imaging*, 12(2):328–333.

Rowe, D. (1999). Bayesian blind source separation. *IEEE Trans. Signal Processing.* submitted.

Saul, L. K., Jaakkola, T., and Jordan, M. I. (1996). Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, 4:61–76.

van der Veen, A.-J. (1997). Analytical method for blind binary signal separation. *IEEE Trans. on Signal Processing*, 45(4):1078–1082.

Initialization: Eqs. (16),(17) and (21)
$\mathbf{J} := \mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{A}$
$\mathbf{h} := \mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{X}$
$\langle\mathbf{S}\rangle := 0$ (or small random values if 0 is a fixed point)
for $m := 1,\ldots,M$ and $t := 1,\ldots,N$:
$\lambda_{mt} := J_{mm}$
endfor
$N_{\langle S\rangle} := 20, N_\lambda := 10, N_{\mathbf{A}} := 10, N_{\mathbf{\Sigma}} := 1$, ftol $:= 10^{-5}$
do:
Expectation-step:
for $N_{\langle S\rangle}$ iterations, eqs. (31) and (22)
for $m := 1,\ldots,M$ and $t := 1,\ldots,N$:
$\gamma_{mt} = h_{mt} - \sum_{m'}(J_{mm'} - \lambda_{m't}\delta_{mm'})\langle S_{m't}\rangle$
$\delta\langle S_{mt}\rangle := f(\gamma_{mt},\lambda_{mt}) - \langle S_{mt}\rangle$
endfor
$\langle\mathbf{S}\rangle := \langle\mathbf{S}\rangle + \eta\delta\langle\mathbf{S}\rangle$
endfor
for N_λ iterations, eqs. (30) and (29)
for $m := 1,\ldots,M$ and $t := 1,\ldots,N$:
$\Lambda_{mt} := \lambda_{mt} + \frac{1}{f'(\gamma_{mt},\lambda_{mt})}$
endfor
for $m := 1,\ldots,M$ and $t := 1,\ldots,N$:
$\delta\lambda_{mt} := \frac{1}{[(\mathbf{\Lambda}_t+\mathbf{J})^{-1}]_{mm}} - \frac{1}{f'(\gamma_{mt},\lambda_{mt})}$
$\lambda_{mt} := \lambda_{mt} + \delta\lambda_{mt}$
endfor
endfor
for $t := 1,\ldots,N$, eq. (27)
$\boldsymbol{\chi}^t := (\mathbf{\Lambda}_t + \mathbf{J})^{-1}$
endfor
Maximization-step
for $N_{\mathbf{A}}$ iterations, eq. (13) or (10)
for $d := 1,\ldots,D$ and $m := 1,\ldots,M$:
$\delta A_{dm} := \frac{[\mathbf{\Sigma}^{-1}\mathbf{X}\langle\mathbf{S}\rangle^T]_{dm}}{[\mathbf{\Sigma}^{-1}\mathbf{A}\langle\mathbf{S}\mathbf{S}^T\rangle]_{dm}+\alpha A_{dm}+\beta} A_{dm} - A_{dm}$
$A_{dm} := A_{dm} + \delta A_{dm}$
endfor
endfor
for $N_{\mathbf{\Sigma}}$ iterations, eq. (9)
$\delta\mathbf{\Sigma} := \frac{1}{N}\langle(\mathbf{X}-\mathbf{A}\mathbf{S})(\mathbf{X}-\mathbf{A}\mathbf{S})^T\rangle - \mathbf{\Sigma}$
$\mathbf{\Sigma} := \mathbf{\Sigma} + \delta\mathbf{\Sigma}$
endfor
$\mathbf{J} := \mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{A}$
$\mathbf{h} := \mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{X}$
while $\max(|\delta\langle S_{mt}\rangle|^2, |\delta\lambda_{mt}|^2, |\delta A_{dm}|^2, |\delta\Sigma_{dd'}|^2) >$ ftol

Table 2: Pseudo-code for the mean field ICA algorithms.