TECHNICAL REPORT

Explaining slow convergence of EM in low noise linear mixtures

Kaare Brandt Petersen and Ole Winther
Informatics and Mathematical Modelling
Technical University of Denmark

January 2005

Contents

1 Introduction

This report conducts an investigation of the convergence properties of the EM algorithm used for linear mixture models. Since the linear mixture model is a rather general approach, the analysis is relevant for a wide range of models which to some degree are subsets of each other: Independent Component Analysis (ICA), probabilistic PCA, Factor Analysis (FA), Independent Factor Analysis (IFA) and Mean Field ICA.

1.1 Timeline on ICA Using EM

In order to present an overview of the results of using the EM algorithm for ICA and related models, is here a short list of selected set of papers each described in a few words.

- 1994 Belouchrani et al. [3]: Seemingly the first paper to suggest the EM algorithm for an ICA problem. They estimate the mixing matrix and the noise covariance. The sources here are assumed discrete QAM4 symbols and the integrals are possible to complete due to the discreteness of the signals.

- 1997 Moulines et al. [8]: Introduction of the EM algorithm for noisy convolutive mixtures with a focus on the instantaneous case. They use MoG as priors to avoid the intractable integrals in the posterior average.

- 1999 Bermond et al. [4]: Update of a square mixing matrix is Taylor expanded in the noise variance. Doing this, they are able to demonstrate several points: (i) In the low noise limit, EM is freezing in the sense that $\mathbf{A}_{n+1} = \mathbf{A}_n$. (ii) To first order in the noise variance, there is no correction for noise compared to Bell and Sejnowski (BS) [2]. That is, the noise model is not making a better estimate than BS when σ^4 is negligible.

- 1999 Attias [1]: A noisy instantaneous mixture model is solved with an EM algorithm using MoG priors. Compared to previous publications he is considering also source reconstruction and situation of having many sources.

- 2000 Lappalainen [7]: In the square case compensating for the slowdown proved earlier by a Taylor expansion in the noise variance, by removal of the "source parts" of the update. They also demonstrates that "Fast ICA" is in fact a special case of this speedup technique.

- 2001 Welling et al. [11]: A so-called constrained EM, where A is a scaling of a rotation, $\mathbf{A} = \alpha\mathbf{R}$. Priors are MoG.

- 2002 Højen-Sørensen et al. [6]: As Moulines an EM algorithm for ICA, but here the major contribution is the use of mean field theory for the posterior means which makes other priors than MoG possible.

- 2002 Deligne et al. [5]: A convolutive ICA using EM. Problems with convergence has been addressed through special initial conditions. MoG Priors.
- 2003 Salakhudinov et al. [10]: The adaptive over-relaxed approach to the problem of slow convergence. Not specific for the ICA using EM, but indeed applicable.
- 2005 Petersen et al. [9]: Small statistical Investigation for MF-ICA. In this they demonstrate the benefit of using the Adaptive EM and a BFGS method.

1.2 The contents of this report

The above list illustrates that many aspects have already been investigated. In the context of this report, especially the paper of Bermond et. al. [4] is central since it has presented many of the central results also to be presented in this report. The reason is that this report is documenting an independent rediscovery of the results. The contribution of this report is

- A detailed derivation.
- A different investigation of the overdetermined case.
- Insight into why Adaptive Overrelaxed EM works.

The structure of the report is as follows: In Sec 2 is described the observation model and the basics of the EM algorithm. Sec 3 is a derivation of the saddle point approximation, while Sec 4 is analyzing the EM algorithm using the saddle point approximation. Finally the conclusions are gathered in sec 5 and computational details are in the appendix.

2 Model and Estimation

2.1 Observation Model

The models under consideration are the linear mixture models, i.e. models of the type

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

where the sources are distributed according to some prior and the noise is zero-mean gaussian, $\boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. The observation vectors $\mathbf{x}_t$ can be collected into a larger matrix $\mathbf{X}$ as columns, such that the dimensions of $\mathbf{X}$ is $M \times N$. Correspondingly we can collect all the source vectors into a larger source matrix $\mathbf{S}$ and write the mixing process including all time steps as $\mathbf{X} = \mathbf{AS} + \boldsymbol{\Gamma}$, where $\boldsymbol{\Gamma}$ is the matrix of the noise. In this more compact notation, $\mathbf{X}$ conveniently denotes the entire data set and we only use the vector notation when ever it is needed to denote a single observation vector. From the assumption of gaussian noise we get a gaussian observation model

$$p(\mathbf{X}|\mathbf{S}) = \frac{1}{\sqrt{|2\pi\mathbf{\Sigma}|}^N} \exp\Big[-\frac{1}{2}\mathrm{Tr}[(\mathbf{X}-\mathbf{AS})^T\mathbf{\Sigma}^{-1}(\mathbf{X}-\mathbf{AS})]\Big]$$

in which the product over all time steps has been expressed as a trace in the exponential function.

2.2 Maximum Likelihood using the EM algorithm

In estimating the parameters $\mathbf{A}$ and $\mathbf{\Sigma}$, a maximum likelihood using the EM algorithm is used. By rewriting $p(\mathbf{X})$ using the hidden variables $\mathbf{S}$ it becomes apparent that we at least in principle have everything we need, knowing the observation model and the source priors. The parameters are estimated as the optimal points of the log likelihood

$$\mathbf{0} = \frac{\partial \ln p(\mathbf{X})}{\partial \mathbf{A}} \qquad \mathbf{0} = \frac{\partial \ln p(\mathbf{X})}{\partial \mathbf{\Sigma}}$$

These equations have the solution

$$\mathbf{A} = \mathbf{X}\langle\mathbf{S}^T\rangle\langle\mathbf{SS}^T\rangle^{-1} \qquad \mathbf{\Sigma} = \langle(\mathbf{X}-\mathbf{AS})(\mathbf{X}-\mathbf{AS})^T\rangle/m$$

where $\langle\cdot\rangle$ denotes average with respect to the source posterior $p(\mathbf{S}|\mathbf{X})$. In cases where $\mathbf{\Sigma}$ is further constrained in some way, it is straight forward to compute the corresponding expression. If we assume that we can compute the first and second moment of the sources with resepct to the source posterior, we now have the two basic steps which is put in to the EM algorithm:

$$\textbf{E-step}: \qquad \langle\mathbf{S}\rangle = \int \mathbf{S}p(\mathbf{S}|\mathbf{X})d\mathbf{S} \qquad \langle\mathbf{SS}^T\rangle = \int \mathbf{SS}^T p(\mathbf{S}|\mathbf{X})d\mathbf{S}$$

$$\textbf{M-step}: \qquad \mathbf{A} = \mathbf{X}\langle\mathbf{S}^T\rangle\langle\mathbf{SS}^T\rangle^{-1} \qquad \mathbf{\Sigma} = \langle(\mathbf{X}-\mathbf{AS})(\mathbf{X}-\mathbf{AS})^T\rangle/m$$

Starting from some reasonable values, we iteratively update our estimate of the source posterior moments and the model parameters, and we are guaranteed in each step an increase of the likelihood.

The source posterior moments are in general not easy to compute. If the source priors are gaussians (FA), mixture of gaussians (IFA) or some other well-chosen family of distribution, then we can compute the moments exactly, but otherwise we must resort to approximations as in MF-ICA, which in many cases are sufficiently accurate to produce good results.

3 Saddle Point Approximation

In this section, as a general approach to the difficulty of computing the source posterior moments, we make an approximation of the integral involved in the posterior and obtain an expansion of the moments in orders of the noise. In this way we are able to study some general properties of the algorithm in the low noise limit.

For simplicity of the equations, we assume for the rest of the paper that the noise is isotropic, $\mathbf{\Sigma} = \sigma^2 \mathbf{I}$. In our experience this does not change any of the results obtained. Furthermore, since the time steps are assumed independent and the log likelihood therefore a sum over over time, we need in this section only to consider one single time step in the saddle point approximation.

3.1 Approximation of an Integral

The technique applied here is the so-called saddle-point approximation of the integral, which is well-known in statistical physics and in large parts of applied mathematics. To lowest order, it is basically an approximation of the integral $I = \int \exp[-f(s)]ds$ which for very large values of $f(s)$ and if $f(s)$ is uni-modal, becomes dominated by its minimum $f(s_0)$. The point s_0 in which the function has its minimum is called the saddle-point, and the approximation is to lowest order $I \cong \exp[-f(s_0)]$. The application of the technique in this paper is slightly more complicated since we have more dimensions, are expanding to higher order, and considering the low noise limit, but the basic idea is the same. Inspired by the equation $\ln p(\mathbf{x}) = \ln \int e^{\ln p(\mathbf{x},\mathbf{s})} d\mathbf{s}$, we define

$$g(\mathbf{s}, \mathbf{h}) = -\sigma^2 \left(\ln p(\mathbf{x}|\mathbf{s}) + \ln p(\mathbf{s}) + \mathbf{h}^T \mathbf{s} \right)$$

Note that for $\mathbf{h} = \mathbf{0}$ the function g is a scaled version of the joint distribution $\ln p(\mathbf{x}, \mathbf{s})$. The somewhat artificial variable $\mathbf{h}$ is a utility variable which will help us obtain the posterior moments as we shall se shortly. Defining the function $g(\mathbf{s}, \mathbf{h})$ as above, we get that

$$\ln I = \ln \int e^{-\frac{1}{\sigma^2} g(\mathbf{s},\mathbf{h})} d\mathbf{s}$$

is a moment generating function in the sense

$$\langle \mathbf{s} \rangle = \frac{\partial \ln I}{\partial \mathbf{h}}\Big|_{\mathbf{h}=\mathbf{0}} \qquad \langle \mathbf{s}\mathbf{s}^T \rangle - \langle \mathbf{s} \rangle \langle \mathbf{s} \rangle^T = \frac{\partial^2 \ln I}{\partial \mathbf{h} \partial \mathbf{h}^T}\Big|_{\mathbf{h}=\mathbf{0}}$$

where $\langle \cdot \rangle$ denotes average with respect to the source posterior $p(\mathbf{s}|\mathbf{x})$. For small σ^2 we can approximate $\ln I$ by a Taylor expansion of $g(\mathbf{s}, \mathbf{h})$ in the saddle point $\mathbf{s}_0$ defined by $\mathbf{g}'(\mathbf{s}_0) = \mathbf{0}$. From this we obtain the saddle point approximation to the integral in $\ln I$

$$\begin{aligned} \ln I &= \ln \int e^{-\frac{1}{\sigma^2} g(\mathbf{s},\mathbf{h})} d\mathbf{s} \\ &= \ln \int e^{-\frac{1}{\sigma^2} g(\mathbf{s}_0,\mathbf{h}) - \frac{1}{2}\frac{1}{\sigma^2}(\mathbf{s}-\mathbf{s}_0)^T \mathbf{g}_0'' (\mathbf{s}-\mathbf{s}_0)} d\mathbf{s} + \mathcal{O}(\sigma^2) \\ &= -\frac{1}{\sigma^2} g(\mathbf{s}_0) + \ln \sqrt{\det[2\pi\sigma^2 (\mathbf{g}_0'')^{-1}]} + \mathcal{O}(\sigma^2) \qquad (1) \end{aligned}$$

In the limit of small σ^2, the derivative of the second term is scaling as $\mathcal{O}(\sigma^4)$ and thus the approximation is efficiently dominated by the first term $-g(\mathbf{s}_0)/\sigma^2$.

3.2 Finding the Saddle Point

In the above we have used the saddle point without explicitly knowing its value. To find it, we make use of the definition $\mathbf{g}'(\mathbf{s}_0) = \mathbf{0}$ but also an approximation of the saddle point in the zero-noise regime, i.e. we write $\mathbf{s}_0$ as a Taylor expansion in σ^2

$$\mathbf{s}_0 = \hat{\mathbf{s}}_0 + \sigma^2 \tilde{\mathbf{s}}_0 + \mathcal{O}(\sigma^4)$$

The zeroth order term $\hat{\mathbf{s}}_0$ is the least squares solution

$$\hat{\mathbf{s}}_0 = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{x}$$

which simplifies further when $\mathbf{A}$ is square. We now want an expression for the first order term $\tilde{\mathbf{s}}_0$ and to find this we insert the combined expression for $\mathbf{s}_0$ into the equation $\mathbf{g}'(\mathbf{s}_0) = \mathbf{0}$ and rearrange the terms to first order in σ^2. Doing that we obtain

$$\tilde{\mathbf{s}}_0 = (\mathbf{A}^T\mathbf{A})^{-1}\left[\frac{\mathbf{p}'(\hat{\mathbf{s}}_0)}{p(\hat{\mathbf{s}}_0)} + \mathbf{h}\right]$$

Having an expression for the saddle point orders of σ^2, makes it possible for us to use the approximation of the integral to obtain the expressions for the posterior moments.

3.3 The Posterior Moments

Inserting into Eq. 1 and performing the calculations, we get

$$\langle \mathbf{s} \rangle = \left.\frac{\partial \ln I}{\partial \mathbf{h}}\right|_{\mathbf{h}=\mathbf{0}} = \hat{\mathbf{s}}_0 + \sigma^2\tilde{\mathbf{s}}_0 + \mathcal{O}(\sigma^4) \quad (2)$$

$$\langle \mathbf{s}\mathbf{s}^T \rangle - \langle \mathbf{s} \rangle\langle \mathbf{s} \rangle^T = \left.\frac{\partial^2 \ln I}{\partial \mathbf{h}\partial \mathbf{h}^T}\right|_{\mathbf{h}=\mathbf{0}} = \sigma^2(\mathbf{A}^T\mathbf{A})^{-1} \quad (3)$$

Note that the posterior mean is the saddle point. This is due to the fact that a minimization of $g(\mathbf{s}, \mathbf{h})$ corresponds to a maximization of $\ln[p(\mathbf{x}, \mathbf{s})]$ which through Bayes theorem has the same maximum as the posterior $p(\mathbf{s}|\mathbf{x})$. In the approximation of the integral we have implicitly assumed that the posterior is uni-modal and symmetric and therefore is the mean the same as the most probable value. The variance provide us with the second moment through

$$\langle \mathbf{s}\mathbf{s}^T \rangle = \hat{\mathbf{s}}\hat{\mathbf{s}}^T + \sigma^2\mathbf{B}_t + \mathcal{O}(\sigma^4)$$

$$\mathbf{B}_t = (\mathbf{A}^T\mathbf{A})^{-1} + (\mathbf{A}^T\mathbf{A})^{-1}\frac{\mathbf{p}'(\hat{\mathbf{s}})}{p(\hat{\mathbf{s}})}\hat{\mathbf{s}}^T + \hat{\mathbf{s}}\frac{\mathbf{p}'(\hat{\mathbf{s}})}{p(\hat{\mathbf{s}})}^T(\mathbf{A}^T\mathbf{A})^{-1}$$

To summarize we now have approximations for the posterior moments in the low noise regime. In order to investigate the quality of this approximation, we have computed the exact and approximated moments for a mixture of gaussians. The result is plotted in Fig. 1. As expected is the approximation not good for large noise variance, but for noise variance in the area of 10^{-2} to 10^{-3} and smaller, the saddle point approximation is very accurate.

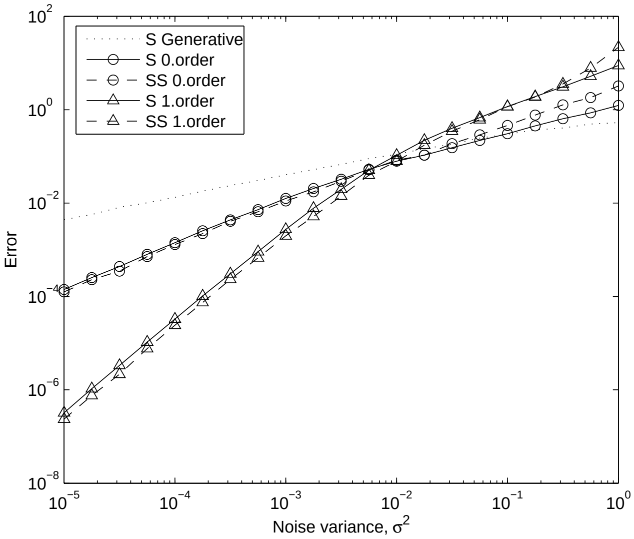

Figure 1: Approximation of the moments - how accurate is it? This plot shows how accurate the saddle point approximation is for different noise levels. The error is defined as the squared mean difference between the approximation and the true posterior mean. For perspective, we have included the generating sequence as well. The prior chosen here is a mixture of two zero-mean gaussians with variance 1 and 1/100.

3.4 The EM Updates

Combining the posterior moments for each time step into larger matrices for the entire data set, we get

$$\begin{aligned} \langle \mathbf{S} \rangle &= \hat{\mathbf{S}} + \sigma^2 \tilde{\mathbf{S}} + \mathcal{O}(\sigma^4) && (4) \\ \langle \mathbf{S}\mathbf{S}^T \rangle &= \hat{\mathbf{S}}\hat{\mathbf{S}}^T + \sigma^2 \mathbf{B} + \mathcal{O}(\sigma^4) && (5) \end{aligned}$$

where $\mathbf{B} = \sum_t \mathbf{B}_t$. From this we can directly insert into Eq. XXX and obtain the EM updates of $\mathbf{A}$ and σ^2

$$\begin{aligned} \mathbf{A}_{n+1} &= \mathbf{X}\hat{\mathbf{S}}(\hat{\mathbf{S}}\hat{\mathbf{S}}^T)^{-1} + \sigma^2 \mathbf{X}\mathbf{F}^T(\hat{\mathbf{S}}\hat{\mathbf{S}}^T)^{-1} + \mathcal{O}(\sigma^4) && (6) \\ \sigma^2_{n+1} &= \sigma^2_{bias} + \sigma^2 \Big(\frac{\mathrm{rank}(\mathbf{A})}{m} - \frac{2}{m}\mathrm{Tr}(\mathbf{U}) \Big) + \mathcal{O}(\sigma^4) && (7) \end{aligned}$$

where $\mathbf{F} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{p}'(\hat{\mathbf{S}})/p(\hat{\mathbf{S}}) - \hat{\mathbf{S}}^T(\hat{\mathbf{S}}\hat{\mathbf{S}}^T)^{-1}\mathbf{B}$ and $\sigma^2_{bias}, \mathbf{U}$ will be explained shortly. These equations are general in the sense that they apply for both square and overdetermined mixing matrix and they simplify considerably in the square case which we analyze in the subsection below.

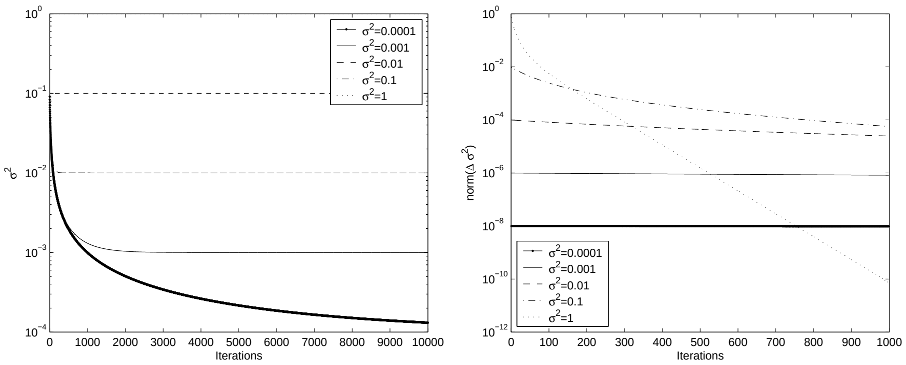

Figure 2: Convergence of the noise for a square FA model: The left plot shows how the noise converges to the levels of the generative model and that lower noise result in slower convergence. The right plot shows that the change of noise variance $\Delta\sigma^2$ is indeed proportional to σ^4. The drop in $\Delta\sigma^2$ seen for $\sigma^2 = 1$ is due to the fact that the model is converged from a very early stage.

4 Analysis

4.1 The Square Case

In the square case we have $\sigma^2_{bias} = 0$ and $\mathbf{U} = \mathbf{0}$, and the expressions simplify into

$$\begin{aligned} \mathbf{A}_{n+1} &= \mathbf{A}_n + \sigma^2 \tilde{\mathbf{A}}_n + \mathcal{O}(\sigma^4) \\ \sigma^2_{n+1} &= \sigma^2_n + \mathcal{O}(\sigma^4) \end{aligned}$$

$$\tilde{\mathbf{A}}_n = \mathbf{A}^{-T} \left(\mathbf{A}^T + \tfrac{1}{N} \hat{\mathbf{Q}} \mathbf{X}^T \right) \mathbf{A}$$

where $\mathbf{Q}_t = \mathbf{p}'(\mathbf{S}_t)/p(\mathbf{S}_t)$ and we have assumed the data are whited, $\mathbf{X}\mathbf{X}^T = N\mathbf{I}$. Many important conclusions can be reached from these equations: (i) the update of the mixing matrix is "freezing" in the sense that to zeroth order the new estimated mixing matrix is identical to the previous one. (ii) The first order correction of the mixing matrix is scaling with the change derived from the noiseless Bell & Sejnowski algorithm (the expression in the parenthesis). This shows that to fist order in the noise variance, there are no improvement of the noisy model compared to the noiseless counterpart. The consequence is that only for noise variances sufficiently large to make σ^4 important, is the noise model different (and hopefully better) than the noiseless model. (iii) The update of the noise is scaling with σ^4 which makes the change in the estimated noise extremely slow to converge when the noise levels are small.

4.2 The Overdetermined Case

To simplify the equations reasonably it is of even greater importance to assume that the data is whitened, i.e. $\mathbf{X}\mathbf{X}^T/N = \mathbf{I}$. When that is the case we have that

$$\mathbf{A}_{n+1} = \mathbf{A}_n + \sigma^2 \tilde{\mathbf{A}}_n + \mathcal{O}(\sigma^4)$$

$$\tilde{\mathbf{A}}_n = (\mathbf{I} - \mathbf{A}\mathbf{A}^+)\frac{\mathbf{X}\hat{\mathbf{Q}}^T}{N} - \mathbf{A} - \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\frac{\hat{\mathbf{Q}}\mathbf{X}^T}{N}\mathbf{A}$$

$$\sigma^2_{bias} = 1 - \frac{\text{rank}(\mathbf{A})}{m}, \qquad \mathbf{U} = \mathbf{\Delta}^T \mathbf{X}\mathbf{Q}^T \hat{\mathbf{A}}^+/N$$

where $\mathbf{\Delta} = \mathbf{I} - \mathbf{A}\mathbf{A}^+$. Following should be noted about these equations: (i) As in the square case, the update in the mixing matrix in freezing, resulting in extremely slow convergence in the low noise limit. (ii) In the correction term for mixing matrix is, we can recognize the structure which leads to the BS solution in the square case and see that it is the first term which is the difference. Not surprisingly, this term is increasing as the ratio of sensors and sources are increasing. (iii) The noise now also have a bias term. On an intuitive level this can be explained as the spill-over from the misjudgement of the number of sources.

4.3 Adaptive Overrelaxed EM

The EM-variant Adaptive Overrelaxed EM is found in [10] and is a very easy and general applicable speedup suggestion. Basically one is enlarging the step size proposed by EM with a factor η

$$\mathbf{A}_{n+1} = \mathbf{A}_n + \eta(\mathbf{A}^{EM}_{n+1} - \mathbf{A}_n)$$

When we combine the low noise analysis with this idea we get

$$\mathbf{A}_{n+1} = \mathbf{A}_n + \eta\sigma^2 \tilde{\mathbf{A}} + \mathcal{O}(\sigma^4)$$

That is, the trick of the adaptive Overrelaxed EM is directly countering the problem of the small noise variance. On the downside we are no longer guaranteed an increase in the log likelihood for each step, but this can be controlled with a test-step modification and the combined algorithm works well on ICA [9].

5 Conclusions

The saddle point approximation and the resulting analysis has demonstrated that

- The linear mixture models have bad convergence properties in the low noise limit, both for square and overdetermined case, and both for the mixing matrix and the noise variance.

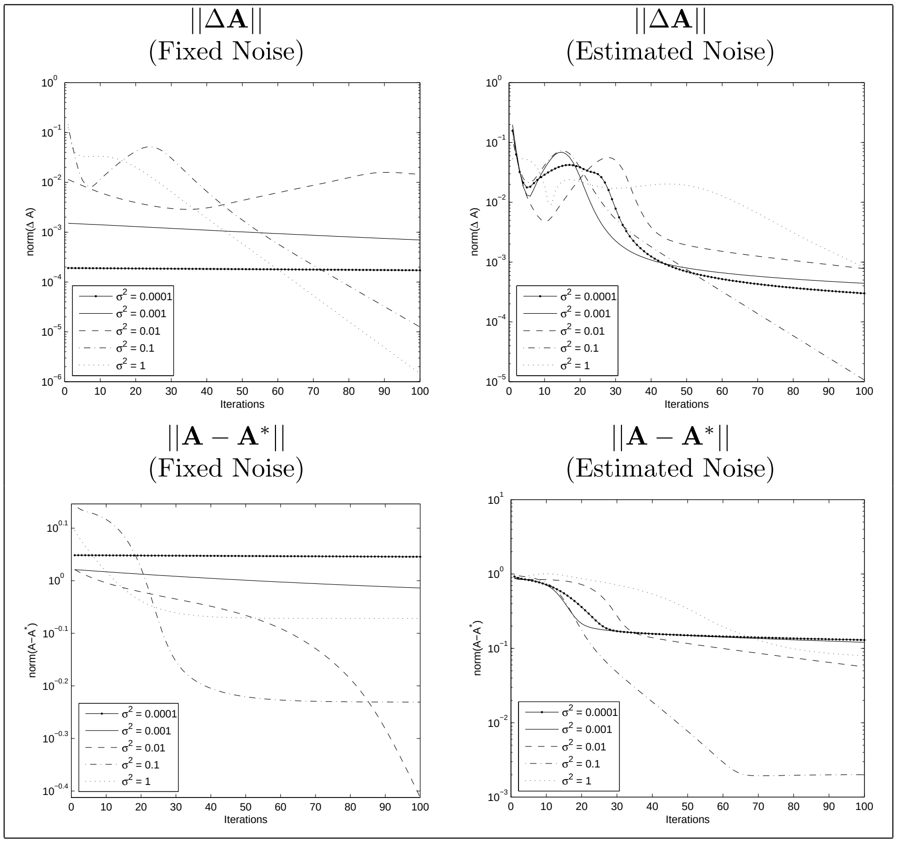

Figure 3: **Plot a)** upper left: The numerical difference in the mixing matrix in a problem with fixed noise at the generative level. Apart from the two largest noise levels, the delta values are exactly where they are predicted. The reason the two largest are not following the predict level is that they are already as converging to whatever solution are optimal for them. **Plot b)** upper right: Same as plot a) but without the fixed noise. Instead the noise is here estimated and results are therefore less clear. The predicted levels are clear obscured by the complications of the combined algorithm in which also initial conditions and choice of mixing matrix makes a difference. **Plot c)** lower left: The difference from the estimated mixing matrix to the optimal. As expected are the low-noise difference very slow to change. **Plot d)** lower right: As plot c) but with estimated noise. Clearly only the example with generative noise level of 1/100 is actually converging to an acceptable solution. This is because lower noise makes the algorithm freeze and higher noise levels make estimation impossible.

- The Overrelaxed Adaptive EM works because the step size directly counters the small noise variance.

Future work, which we are presently working on, is the generalize the presented results to non-linear observation models and to the Variational Bayes EM variant.

A Calculations and Details

In this section we make extensive use of the possible prewhitening of the data $\mathbf{X}\mathbf{X}^T = N\mathbf{I}$ and of the notation

$$\begin{aligned}
\boldsymbol{\alpha} &= \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T \\
&= \mathbf{A}\mathbf{A}^+ \\
\boldsymbol{\Delta} &= \mathbf{I} - \mathbf{A}\mathbf{A}^+ \\
\hat{\mathbf{Q}} &= \mathbf{p}'(\hat{\mathbf{S}})/p(\hat{\mathbf{S}})
\end{aligned}$$

A.1 The Moments

The moments, derived in the text, are repeated here for consistency

$$\begin{aligned}
\langle \mathbf{S} \rangle &= \hat{\mathbf{S}} + \sigma^2 \tilde{\mathbf{S}} + \mathcal{O}(\sigma^4) \\
\hat{\mathbf{S}} &= (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{X} \\
\tilde{\mathbf{S}} &= (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{Q} \\
\langle \mathbf{S}\mathbf{S}^T \rangle &= \hat{\mathbf{S}}\hat{\mathbf{S}}^T + \sigma^2 \mathbf{B} + \mathcal{O}(\sigma^4) \\
\hat{\mathbf{S}}\hat{\mathbf{S}}^T &= (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{X}\mathbf{X}^T\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1} \\
&= N(\mathbf{A}^T\mathbf{A})^{-1} \qquad \text{(whitening)} \\
\mathbf{B} &= N(\mathbf{A}^T\mathbf{A})^{-1} + (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{Q}\hat{\mathbf{S}}^T + \hat{\mathbf{S}}\mathbf{Q}^T(\mathbf{A}^T\mathbf{A})^{-1} \\
&= N(\mathbf{A}^T\mathbf{A})^{-1} + (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{Q}\mathbf{X}^T\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1} + (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{X}\mathbf{Q}^T(\mathbf{A}^T\mathbf{A})^{-1} \\
&= (\mathbf{A}^T\mathbf{A})^{-1}[N\mathbf{A}^T\mathbf{A} + \mathbf{Q}\mathbf{X}^T\mathbf{A} + \mathbf{A}^T\mathbf{X}\mathbf{Q}^T](\mathbf{A}^T\mathbf{A})^{-1}
\end{aligned}$$

A.2 General (Non-square) identities

The following identities turn out to be useful in later calculations

$$\begin{aligned}
\mathbf{X}^T\mathbf{A}\hat{\mathbf{S}} &= \mathbf{X}^T\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{X} \\
\mathbf{X}^T\mathbf{A}\tilde{\mathbf{S}} &= \mathbf{X}^T\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{Q} \\
\hat{\mathbf{S}}\hat{\mathbf{S}}^T &= (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{X}\mathbf{X}^T\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1} \\
\mathbf{A}^T\mathbf{A}\hat{\mathbf{S}}\hat{\mathbf{S}}^T &= \mathbf{A}^T\mathbf{X}\mathbf{X}^T\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1} \\
\mathbf{A}^T\mathbf{A}\mathbf{B} &= N\mathbf{I} + \mathbf{Q}\mathbf{X}^T\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1} + \mathbf{A}^T\mathbf{X}\mathbf{Q}^T(\mathbf{A}^T\mathbf{A})^{-1}
\end{aligned}$$

A.3 The Mixing Matrix

$$
\begin{aligned}
\mathbf{A} &= \mathbf{X}\langle\mathbf{S}\rangle^T\langle\mathbf{S}\mathbf{S}^T\rangle^{-1} \\
&= \mathbf{X}(\hat{\mathbf{S}}+\sigma^2\tilde{\mathbf{S}})^T(\hat{\mathbf{S}}\hat{\mathbf{S}}^T+\sigma^2\mathbf{B})^{-1} \\
&= \mathbf{X}(\hat{\mathbf{S}}+\sigma^2\tilde{\mathbf{S}})^T((\hat{\mathbf{S}}\hat{\mathbf{S}}^T)^{-1}-\sigma^2(\hat{\mathbf{S}}\hat{\mathbf{S}}^T)^{-1}\mathbf{B}(\hat{\mathbf{S}}\hat{\mathbf{S}}^T)^{-1})+\mathcal{O}(\sigma^4) \\
&= \mathbf{X}\hat{\mathbf{S}}^T(\hat{\mathbf{S}}\hat{\mathbf{S}}^T)^{-1}+\sigma^2(\mathbf{X}\tilde{\mathbf{S}}^T(\hat{\mathbf{S}}\hat{\mathbf{S}}^T)^{-1}-\mathbf{X}\hat{\mathbf{S}}^T(\hat{\mathbf{S}}\hat{\mathbf{S}}^T)^{-1}\mathbf{B}(\hat{\mathbf{S}}\hat{\mathbf{S}}^T)^{-1}) \\
&= \mathbf{X}\hat{\mathbf{S}}^T(\hat{\mathbf{S}}\hat{\mathbf{S}}^T)^{-1}+\sigma^2\mathbf{X}(\tilde{\mathbf{S}}^T-\hat{\mathbf{S}}^T(\hat{\mathbf{S}}\hat{\mathbf{S}}^T)^{-1}\mathbf{B})(\hat{\mathbf{S}}\hat{\mathbf{S}}^T)^{-1} \\
&= \mathbf{X}\hat{\mathbf{S}}^T(\hat{\mathbf{S}}\hat{\mathbf{S}}^T)^{-1}+\sigma^2\mathbf{X}\mathbf{F}^T(\hat{\mathbf{S}}\hat{\mathbf{S}}^T)^{-1}
\end{aligned}
$$

Overdetermined: Using $\boldsymbol{\alpha}_x = \mathbf{A}^T\mathbf{X}\mathbf{X}^T\mathbf{A}$

$$
\begin{aligned}
\hat{\mathbf{A}}_n &= \mathbf{X}\mathbf{X}^T\mathbf{A}(\mathbf{A}^T\mathbf{X}\mathbf{X}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{A} \\
\tilde{\mathbf{A}}_n &= \left\{\mathbf{X}\mathbf{Q}^T-\mathbf{X}\mathbf{X}^T\mathbf{A}\boldsymbol{\alpha}_x^{-1}[N\mathbf{A}^T\mathbf{A}+\mathbf{Q}\mathbf{X}^T\mathbf{A}+\mathbf{A}^T\mathbf{X}\mathbf{Q}^T]\right\}\boldsymbol{\alpha}_x^{-1}\mathbf{A}^T\mathbf{A}
\end{aligned}
$$

A.4 The Noise Variance

$$
\begin{aligned}
& \langle(\mathbf{X}-\mathbf{A}\mathbf{S})(\mathbf{X}-\mathbf{A}\mathbf{S})^T\rangle \\
= {} & \mathbf{X}\mathbf{X}^T+\mathbf{A}\langle\mathbf{S}\mathbf{S}^T\rangle\mathbf{A}^T-\mathbf{X}\langle\mathbf{S}\rangle^T\mathbf{A}^T-\mathbf{A}\langle\mathbf{S}\rangle\mathbf{X}^T \\
= {} & \mathbf{X}\mathbf{X}^T+\mathbf{A}(\hat{\mathbf{S}}\hat{\mathbf{S}}^T+\sigma^2\mathbf{B})\mathbf{A}^T-\mathbf{X}(\hat{\mathbf{S}}+\sigma^2\tilde{\mathbf{S}})^T\mathbf{A}^T-\mathbf{A}(\hat{\mathbf{S}}+\sigma^2\tilde{\mathbf{S}})\mathbf{X}^T \\
= {} & \mathbf{X}\mathbf{X}^T+\mathbf{A}\hat{\mathbf{S}}\hat{\mathbf{S}}^T\mathbf{A}^T-\mathbf{X}\hat{\mathbf{S}}^T\mathbf{A}^T-\mathbf{A}\hat{\mathbf{S}}\mathbf{X} \\
& \quad +\sigma^2\left(\mathbf{A}\mathbf{B}\mathbf{A}^T-\mathbf{X}\tilde{\mathbf{S}}^T\mathbf{A}^T-\mathbf{A}\tilde{\mathbf{S}}\mathbf{X}^T\right) \\
= {} & \mathbf{X}\mathbf{X}^T+\mathbf{A}\hat{\mathbf{S}}\hat{\mathbf{S}}^T\mathbf{A}^T-\mathbf{X}\hat{\mathbf{S}}^T\mathbf{A}^T-\mathbf{A}\hat{\mathbf{S}}\mathbf{X} \\
& \quad +\sigma^2\left(\hat{\mathbf{A}}\mathbf{B}\hat{\mathbf{A}^T}-\mathbf{X}\tilde{\mathbf{S}}^T\hat{\mathbf{A}}^T-\hat{\mathbf{A}}\tilde{\mathbf{S}}\mathbf{X}^T\right)+\mathcal{O}(\sigma^4) \\
= {} & (\mathbf{X}-\mathbf{A}\hat{\mathbf{S}})(\mathbf{X}-\mathbf{A}\hat{\mathbf{S}})^T+\sigma^2\left(\ldots_1\right)+\mathcal{O}(\sigma^4) \\
= {} & (\mathbf{X}-(\hat{\mathbf{A}}+\sigma^2\tilde{\mathbf{A}})\hat{\mathbf{S}})(\mathbf{X}-(\hat{\mathbf{A}}+\sigma^2\tilde{\mathbf{A}})\hat{\mathbf{S}})^T+\sigma^2\left(\ldots\right)+\mathcal{O}(\sigma^4) \\
= {} & ((\mathbf{X}-\hat{\mathbf{A}}\hat{\mathbf{S}})-\sigma^2\tilde{\mathbf{A}}\hat{\mathbf{S}})((\mathbf{X}-\hat{\mathbf{A}}\hat{\mathbf{S}})-\sigma^2\tilde{\mathbf{A}}\hat{\mathbf{S}})^T+\sigma^2\left(\ldots_1\right)+\mathcal{O}(\sigma^4) \\
= {} & (\mathbf{X}-\hat{\mathbf{A}}\hat{\mathbf{S}})(\mathbf{X}-\hat{\mathbf{A}}\hat{\mathbf{S}})^T-\sigma^2[(\mathbf{X}-\hat{\mathbf{A}}\hat{\mathbf{S}})\hat{\mathbf{S}}^T\tilde{\mathbf{A}}^T+\tilde{\mathbf{A}}\hat{\mathbf{S}}(\mathbf{X}-\hat{\mathbf{A}}\hat{\mathbf{S}})^T]+\sigma^2\left(\ldots_1\right)+\mathcal{O}(\sigma^4) \\
= {} & (\mathbf{X}-\hat{\mathbf{A}}\hat{\mathbf{S}})(\mathbf{X}-\hat{\mathbf{A}}\hat{\mathbf{S}})^T-\sigma^2[\ldots_2]+\sigma^2\left(\ldots_1\right)+\mathcal{O}(\sigma^4) \\
= {} & (\mathbf{X}-\hat{\mathbf{A}}\hat{\mathbf{S}})(\mathbf{X}-\hat{\mathbf{A}}\hat{\mathbf{S}})^T+\sigma^2\left(\ldots_1-\ldots_2\right)+\mathcal{O}(\sigma^4) \\
= {} & (\mathbf{X}-\hat{\mathbf{A}}(\hat{\mathbf{A}}^T\hat{\mathbf{A}})^{-1}\hat{\mathbf{A}}^T\mathbf{X})(\mathbf{X}-\hat{\mathbf{A}}(\hat{\mathbf{A}}^T\hat{\mathbf{A}})^{-1}\hat{\mathbf{A}}^T\mathbf{X})^T+\sigma^2\left(\ldots_1-\ldots_2\right)+\mathcal{O}(\sigma^4) \\
= {} & (\mathbf{I}-\boldsymbol{\alpha})\mathbf{X}\mathbf{X}^T(\mathbf{I}-\boldsymbol{\alpha})^T+\sigma^2\left(\ldots_1-\ldots_2\right)+\mathcal{O}(\sigma^4) \\
= {} & N(\mathbf{I}-\boldsymbol{\alpha})(\mathbf{I}-\boldsymbol{\alpha})^T+\sigma^2\left(\ldots_1-\ldots_2\right)+\mathcal{O}(\sigma^4) \\
= {} & N\boldsymbol{\Delta}\boldsymbol{\Delta}^T+\sigma^2\left(\ldots_1-\ldots_2\right)+\mathcal{O}(\sigma^4)
\end{aligned}
$$

$$\begin{aligned}
(\ldots_1) &= \hat{\mathbf{A}}\mathbf{B}\hat{\mathbf{A}}^T - \mathbf{X}\tilde{\mathbf{S}}^T\hat{\mathbf{A}}^T - \hat{\mathbf{A}}\tilde{\mathbf{S}}\mathbf{X}^T \\
&= \hat{\mathbf{A}}\Big[(\hat{\mathbf{A}}^T\hat{\mathbf{A}})^{-1}[N\hat{\mathbf{A}}^T\hat{\mathbf{A}} + \mathbf{Q}\mathbf{X}^T\hat{\mathbf{A}} + \hat{\mathbf{A}}^T\mathbf{X}\mathbf{Q}^T](\hat{\mathbf{A}}^T\hat{\mathbf{A}})^{-1}\Big]\hat{\mathbf{A}}^T \\
&\quad -\mathbf{X}((\hat{\mathbf{A}}^T\hat{\mathbf{A}})^{-1}\mathbf{Q})^T\hat{\mathbf{A}}^T - \hat{\mathbf{A}}(\hat{\mathbf{A}}^T\hat{\mathbf{A}})^{-1}\mathbf{Q}\mathbf{X}^T \\
&= (\hat{\mathbf{A}}^+)^T[N\hat{\mathbf{A}}^T\hat{\mathbf{A}} + \mathbf{Q}\mathbf{X}^T\hat{\mathbf{A}} + \hat{\mathbf{A}}^T\mathbf{X}\mathbf{Q}^T]\hat{\mathbf{A}}^+ - \mathbf{X}\mathbf{Q}^T\hat{\mathbf{A}}^+ - (\hat{\mathbf{A}}^+)^T\mathbf{Q}\mathbf{X}^T \\
&= N\hat{\mathbf{A}}\hat{\mathbf{A}}^+ + (\hat{\mathbf{A}}^+)^T\mathbf{Q}\mathbf{X}^T\hat{\mathbf{A}}\hat{\mathbf{A}}^+ + (\hat{\mathbf{A}}\hat{\mathbf{A}}^+)^T\mathbf{X}\mathbf{Q}^T\hat{\mathbf{A}}^+ - \mathbf{X}\mathbf{Q}^T\hat{\mathbf{A}}^+ - (\hat{\mathbf{A}}^+)^T\mathbf{Q}\mathbf{X}^T \\
&= N\boldsymbol{\alpha} + (\hat{\mathbf{A}}^+)^T\mathbf{Q}\mathbf{X}^T(\hat{\mathbf{A}}\hat{\mathbf{A}}^+ - \mathbf{I}) + ((\hat{\mathbf{A}}\hat{\mathbf{A}}^+)^T - \mathbf{I})\mathbf{X}\mathbf{Q}^T\hat{\mathbf{A}}^+ \\
&= N\boldsymbol{\alpha} - \{(\hat{\mathbf{A}}^+)^T\mathbf{Q}\mathbf{X}^T\boldsymbol{\Delta} + \boldsymbol{\Delta}^T\mathbf{X}\mathbf{Q}^T\hat{\mathbf{A}}^+\}
\end{aligned}$$

$$\begin{aligned}
(\ldots_2) &= (\mathbf{X} - \hat{\mathbf{A}}\hat{\mathbf{S}})\hat{\mathbf{S}}^T\tilde{\mathbf{A}}^T + \tilde{\mathbf{A}}\hat{\mathbf{S}}(\mathbf{X} - \hat{\mathbf{A}}\hat{\mathbf{S}})^T \\
&= (\mathbf{X}\hat{\mathbf{S}}^T - \hat{\mathbf{A}}\hat{\mathbf{S}}\hat{\mathbf{S}}^T)\tilde{\mathbf{A}}^T + \tilde{\mathbf{A}}(\mathbf{X}\hat{\mathbf{S}}^T - \hat{\mathbf{A}}\hat{\mathbf{S}}\hat{\mathbf{S}}^T) \\
&= (N\hat{\mathbf{A}}(\hat{\mathbf{A}}^T\hat{\mathbf{A}})^{-1} - N\hat{\mathbf{A}}(\hat{\mathbf{A}}^T\hat{\mathbf{A}})^{-1})\tilde{\mathbf{A}}^T \\
&\qquad +\tilde{\mathbf{A}}(N(\hat{\mathbf{A}}^T\hat{\mathbf{A}})^{-1}\hat{\mathbf{A}}^T - N(\hat{\mathbf{A}}^T\hat{\mathbf{A}})^{-1}\hat{\mathbf{A}}^T) \\
&= \mathbf{0}
\end{aligned}$$

In the square case, $\boldsymbol{\Delta} = \mathbf{0}$ and $\boldsymbol{\alpha} = \mathbf{I}$, and therefore

$$\begin{aligned}
\sigma^2_{n+1} &= \mathrm{Tr}(\langle(\mathbf{X} - \mathbf{A}\mathbf{S})(\mathbf{X} - \mathbf{A}\mathbf{S})^T\rangle)/(Nm) \\
&= \sigma^2 + \mathcal{O}(\sigma^4)
\end{aligned}$$

In the overdetermined case we have

$$\begin{aligned}
\sigma^2_{n+1} &= \mathrm{Tr}(\langle(\mathbf{X} - \mathbf{A}\mathbf{S})(\mathbf{X} - \mathbf{A}\mathbf{S})^T\rangle)/(Nm) \\
&= \mathrm{Tr}(\boldsymbol{\Delta}\boldsymbol{\Delta}^T)/m + \sigma^2\Big(\mathrm{Tr}(\boldsymbol{\alpha})/m - \mathrm{Tr}(\mathbf{U}^T + \mathbf{U})/m\Big) + \mathcal{O}(\sigma^4) \\
&\qquad \text{where } \mathbf{U} = \boldsymbol{\Delta}^T\mathbf{X}\mathbf{Q}^T\hat{\mathbf{A}}^+/N \\
&= \mathrm{Tr}((\mathbf{I} - \boldsymbol{\alpha})(\mathbf{I} - \boldsymbol{\alpha})^T)/m + \sigma^2\Big(\mathrm{Tr}(\boldsymbol{\alpha})/m - \mathrm{Tr}(\mathbf{U}^T + \mathbf{U})/m\Big) + \mathcal{O}(\sigma^4) \\
&= \mathrm{Tr}(\mathbf{I} - 2\boldsymbol{\alpha} + \boldsymbol{\alpha}\boldsymbol{\alpha})/m + \sigma^2\Big(\mathrm{Tr}(\boldsymbol{\alpha})/m - 2\mathrm{Tr}(\mathbf{U})/m\Big) + \mathcal{O}(\sigma^4) \\
&= \mathrm{Tr}(\mathbf{I} - \boldsymbol{\alpha})/m + \sigma^2\Big(\mathrm{Tr}(\boldsymbol{\alpha})/m - 2\mathrm{Tr}(\mathbf{U})/m\Big) + \mathcal{O}(\sigma^4) \\
&= 1 - \frac{\mathrm{Tr}(\boldsymbol{\alpha})}{m} + \sigma^2\Big(\mathrm{Tr}(\boldsymbol{\alpha})/m - 2\mathrm{Tr}(\mathbf{U})/m\Big) + \mathcal{O}(\sigma^4) \\
&= 1 - \frac{\mathrm{rank}(\boldsymbol{\alpha})}{m} + \sigma^2\Big(\frac{\mathrm{rank}(\boldsymbol{\alpha})}{m} - 2\mathrm{Tr}(\mathbf{U})/m\Big) + \mathcal{O}(\sigma^4) \\
&= 1 - \frac{\mathrm{rank}(\mathbf{A})}{m} + \sigma^2\Big(\frac{\mathrm{rank}(\mathbf{A})}{m} - 2\mathrm{Tr}(\mathbf{U})/m\Big) + \mathcal{O}(\sigma^4) \\
&= \sigma^2_{bias} + \sigma^2\Big(\tfrac{\mathrm{rank}(\mathbf{A})}{m} - \tfrac{2}{m}\mathrm{Tr}(\mathbf{U})\Big) + \mathcal{O}(\sigma^4)
\end{aligned}$$

References

[1] Hagai Attias. Independent factor analysis. *Neural Computation*, (11):803–851, 1999.

[2] Anthony J. Bell and Terrence J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

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

[4] O. Bermond and Jean Francois Cardoso. Approximate likelihood for noisy mixtures. In *Proceedings of the ICA Conference*, 1999.

[5] Sabine Deligne and Ramesh A. Gopinath. An em algorithm for convolutive independent component analysis. *Neurocomputing*, 49(1-4):187–211, 2002.

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

[7] Harry Lappalainen and Petteri Pajunen. Fast algorithms for bayesian independent component analysis. In *Proceedings of the ICA Conference*, 2000.

[8] 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.

[9] Kaare Brandt Petersen and Ole Winther. The em algorithm in independent component analysis. In *International Conference on Acoustics, Speech, and Signal Processing*, 2005.

[10] R. Salakhutdinov and S. Roweis. Adaptive overrelaxed bound optimization methods. International Conference on Machine Learning, ICML, 2003.

[11] Max Welling and Markus Weber. A constrained em algorithm for independent component analysis. *Neural Computation*, 13, 2001.