

On the slow convergence of EM and VBEM in low noise linear models

Kaare Brandt Petersen, Ole Winther and Lars Kai Hansen

Contact: kbp@imm.dtu.dk

March 1, 2005

Abstract

We analyze convergence of the expectation maximization (EM) and variational Bayes EM schemes for parameter estimation in noisy linear models. The analysis shows that both schemes are inefficient in the low noise limit. The linear model with additive noise includes as special cases Independent Component Analysis (ICA), probabilistic PCA, Factor Analysis, and Kalman Filtering, hence, the results are relevant for many practical applications.

1 Introduction

The expectation maximization (EM) algorithm introduced by Dempster et al. (Dempster et al., 1977) is widely used for maximum likelihood estimation in hidden variable models. More recently a generalization of the EM algorithm, the so-called variational Bayes EM algorithm (VBEM), has been introduced, see e.g., (Attias, 1999) which allows for more accurate modelling of parameter uncertainty. EM convergence is known to slow down dramatically when the signal-to-noise ratio is high, and a natural question is then: Will the more accurate modelling of parameter variance in VBEM assist the convergence? Here we analyze both schemes and show that they are subject to slow convergence in the low noise limit.

We consider linear models with additive normal noise

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

where $\mathbf{x}_t \in \mathbb{R}^m$ are N observed data vectors, $\mathbf{s}_t \in \mathbb{R}^d$ unobserved, hidden variables and $\mathbf{n}_t \sim \mathcal{N}(\mathbf{0}, \Sigma)$ white gaussian noise. For notational convenience we construct the matrices \mathbf{X} and \mathbf{S} which consists of the observed and unobserved data vectors as columns. The unobserved variables \mathbf{S} are assumed distributed according to a prior $p(\mathbf{S})$, which can be Gaussian (Factor analysis) or non-Gaussian (ICA). The matrix \mathbf{A} is referred to as the mixing matrix and it can be square ($m = d$) but does not have to. If $m < d$ we speak of the *overcomplete* case, while the opposite situation $d < m$ is denoted *overdetermined*. In our discussion, the data is assumed pre-whitened, i.e. $\mathbf{X}\mathbf{X}^T = N\mathbf{I}$, a mild condition that simplifies the notation.

2 Slow convergence in EM

For parameter estimation (\mathbf{A}, Σ) in the linear model, the main challenge is that the marginal likelihood involves an average over all possible configurations of the hidden variables, with a measure that depends on the unknown parameters themselves. EM algorithms break up this stalemate in two separate iterated steps. First we find the posterior distribution of the hidden variables $P(\mathbf{S}|\mathbf{X}, \mathbf{A}, \Sigma)$, for fixed parameters, secondly, we improve the parameters by maximizing the log likelihood averaged w.r.t. the approximate hidden variable posterior, for details consult (Dempster et al., 1977; McLachlan and Krishnan, 1997). Bermond and Cardoso made an important but seemingly little known discovery about the convergence properties of the EM algorithm in the low noise limit (Bermond and Cardoso, 1999). Following their line of thought, and for simplicity considering the case $\Sigma = \sigma^2 \mathbf{I}$, we can expand the moments of the posterior $\langle \mathbf{S} \rangle$ and $\langle \mathbf{S}\mathbf{S}^T \rangle$ in powers of the noise variance (see Fig. 1) to obtain approximate expressions for the parameter updates. Using the notation $\mathbf{\Gamma} = \mathbf{X} - \mathbf{A}\mathbf{S}$ we get

$$\begin{aligned} \mathbf{A}_{n+1} &= \mathbf{X}\langle \mathbf{S} \rangle^T \langle \mathbf{S}\mathbf{S}^T \rangle^{-1} = \mathbf{A}_n + \sigma_n^2 \tilde{\mathbf{A}}_n + \mathcal{O}(\sigma^4) \\ \sigma_{n+1}^2 &= \frac{1}{mN} \text{Tr}(\langle \mathbf{\Gamma}\mathbf{\Gamma}^T \rangle) = \sigma_{bias}^2 + \sigma_n^2 z + \mathcal{O}(\sigma^4) \end{aligned}$$

where \mathbf{A}_n denotes the estimated mixing matrix in the n .th iteration. In the square case, the noise update simplifies into $\sigma_{n+1}^2 = \sigma_n^2 + \mathcal{O}(\sigma^4)$. In the overdetermined case, $\sigma_{bias}^2 = 1 - \text{rank}(\mathbf{A})/m$ and $z = \text{rank}(\mathbf{A})/m - 2\text{Tr}(\mathbf{U})/N$, where \mathbf{U} is a data and prior dependent matrix. This result is discussed in (Petersen and Winther, 2005a) to which the reader is referred for details.

The result indeed explains the poor convergence properties experienced using EM in

the low noise limit. The EM algorithm ‘freezes’ and an excessive amount of iterations is needed for convergence of the mixing matrix. Moreover for the square case, as also mentioned in (Bermond and Cardoso, 1999), the first order correction $\tilde{\mathbf{A}}_n$ is proportional to the gradient of the noiseless model’s likelihood and thus the fix point is to first order equivalent to the fix point of the noiseless model (Bell and Sejnowski, 1995).

The slow convergence of the EM algorithm has been debated for a while and many suggestions for speeding up the EM algorithm have been proposed (McLachlan and Krishnan, 1997). One straight forward method is to use a gradient based optimizer in the M-step. The gradient and the bound value are expressed in terms of the sufficient statistics which are obtained in the E-step (Olsson et al., 2005). Recently, another general technique by Salakhudinov et. al. (Salakhutdinov and Roweis, 2003), called *adaptive over-relaxed EM* was proposed leading to considerable faster convergence (Petersen and Winther, 2005b). The key idea of the adaptive over-relaxed EM is to boost the update by a factor $\eta \geq 1$. Combining this with the low noise limit analysis we get

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

That is, adaptive over-relaxed EM works because the step size factor η directly counters the small magnitude of the noise variance. The only downside is that there are no longer a guarantee of an increase in the likelihood and a ”test-step” is introduced to remedy this.

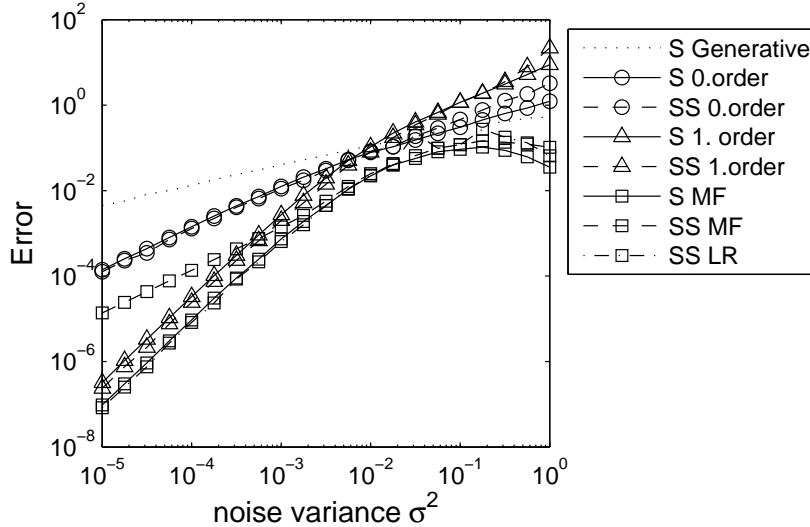


FIGURE 1: This plot demonstrates that the fundamental Taylor expansion of the moments $\langle \mathbf{S} \rangle$ and $\langle \mathbf{S}\mathbf{S}^T \rangle$ is reasonably accurate. A set of sources \mathbf{S}_{gen} is generated using a mixture of gaussians (MoG) prior and from this, using a suitable 2x2 mixing matrix, the observed data \mathbf{X} is constructed for each noise level. Since the source prior is a MoG, the exact posterior moments $\langle \mathbf{S} \rangle_{exc}$, $\langle \mathbf{S}\mathbf{S}^T \rangle_{exc}$ can be computed. The error is the mean squared difference of the true mean and the approximation, $Err = \frac{1}{dN} \sum_{it} (\langle \mathbf{S}_{it} \rangle_{exc} - \langle \mathbf{S}_{it} \rangle_{est})^2$ and correspondingly for the second moment. Note that: 1) The approximation is fairly accurate when the noise variance is small. 2) As expected, the first order approximation (triangles) is more accurate than the zero.th order approximation (circles) in the low noise regime. 3) Only for noise variance larger than 10^{-2} is it beneficial to use the generative sources (dots) as estimators for the posterior means. This is of course only possible for artificial data sets, but included for perspective. 4) The Mean Field (MF) approximations to the posterior moments (squares) is also included for perspective, see (Hojen-Sorensen et al., 2002) for details. The MF approach is performing very well indeed, especially when the so-called linear response (LR) correction is taken into account. This is an indicator that in the low noise regime, ICA techniques such as Mean Field ICA may prove to be accurate approaches.

3 Variational Bayes EM

In variational Bayes EM we expand the model to include a distribution over the model parameters \mathbf{A} and σ^2 , i.e., treating them at the same footing as the hidden variables. See, e.g., (Beal and Ghahramani, 2003) for an introduction to variational Bayes techniques. The algorithm is aimed at maximizing the lower bound of the marginal log-likelihood and allows for convenient addition of prior information on the parameters. We choose a zero-mean gaussian prior for the mixing matrix, with covariance Σ_p , and an inverse gamma distribution with (hyper) parameters α_p and β_p .

$$p(\mathbf{A}) \propto \exp[-\frac{1}{2}\text{Tr}(\mathbf{A}\Sigma_p^{-1}\mathbf{A}^T)]$$

$$p(\sigma^2) \propto (\sigma^2)^{-(\alpha_p+1)} \exp[-\beta_p/\sigma^2]$$

Combining these priors with the observation model we obtain the variational approximations for the posterior distributions, which have the moments that are updated sequentially in the VBEM algorithm. At convergence we use the posterior mean of these variational distributions as estimators of the unknown model parameters.

The statistics that determines the width of the posterior distribution of \mathbf{A} is $\langle 1/\sigma^2 \rangle$. Defining $r^2 = 1/\langle 1/\sigma^2 \rangle$ the low noise limit corresponds to $r^2 \rightarrow 0$ and we can expand the the moments involved in update of the \mathbf{A} pdf, in powers of r^2

$$\langle \mathbf{A} \rangle_{n+1} = \mathbf{X} \langle \mathbf{S} \rangle^T [\langle \mathbf{S}\mathbf{S}^T \rangle + r^2 \Sigma_p^{-1}]^{-1} = \langle \mathbf{A} \rangle_n + \mathcal{O}(r^2)$$

$$\text{Var}(\mathbf{A})_{n+1} = r^2 [\langle \mathbf{S}\mathbf{S}^T \rangle + r^2 \Sigma_p^{-1}]^{-1} = \mathbf{0} + \mathcal{O}(r^2)$$

and accordingly for the parameters α, β of the inversed-gamma distributed σ^2

$$\begin{aligned}\alpha_{n+1} &= \alpha_p + \frac{mN}{2} &= \alpha_n \\ \beta_{n+1} &= \beta_p + N\beta_{n+1}^{bias} &= \beta_n + N\mathcal{O}(r^2) \\ r_{n+1}^2 &= \beta_{n+1}/\alpha_{n+1} &= r_n^2 + \mathcal{O}(r^2)\end{aligned}$$

where $\beta_{n+1}^{bias} = 1 - \text{rank}(\langle \mathbf{A} \rangle_{n+1})/m$. Hence, we find that the VBEM update for the mixing matrix and the crucial moment of the noise distribution is 'freezing' exactly as in EM.

4 Discussion

The analysis shows that for linear models with low Gaussian noise, *both* the traditional EM algorithm and the variational Bayes extension, which practically degenerates back into an EM algorithm, have serious defects with respect to the rate of convergence. Experience from ICA problems furthermore indicates that the window in which the noise is sufficiently large to make the convergence reasonable and yet not too large with respect to estimation of parameters, is indeed very small.

Furthermore note, that in (Salakhutdinov et al., 2003) the convergence rate in a gaussian mixture model is demonstrated to be slow when the noise level is large, i.e. when the mixtures have considerable overlap. The situation analyzed in this paper, however, is a limit of low noise in which the problem intuitively should have an extraordinary clear and well-defined solution. In that sense this result is counter-intuitive and different from some of the previous observations regarding the slowdown of the EM algorithm. Most likely the explanation is, that there is more than one situation

in which the EM algorithm becomes slow and that these different situations are not effects of the same underlying reason, but rather truly different.

Finally, it is of course crucial for the analysis that the observation model is linear, since we otherwise cannot get closed form expressions in the M-step. But practical experience and preliminary analysis suggests that this it is not the core of convergence problem and we are instead conjecturing that it is indeed the low noise limit which is the essence of the matter.

Acknowledgements

The research for this publication was supported financially by Oticon Fonden.

References

- Attias, H. (1999). Inferring parameters and structure of latent variable models by variational bayes. In *In Proceedings of Fifteenth Conference on Uncertainty in Artificial Intelligence, UAI*.
- Beal, M. J. and Ghahramani, Z. (2003). The Variational Bayesian EM Algorithm for Incomplete Data: With Application to Scoring Graphical Model Structures. *Bayesian Statistics*, (7).
- Bell, A. J. and Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159.

- Bermond, O. and Cardoso, J. F. (1999). Approximate likelihood for noisy mixtures. In *Proceedings of the ICA Conference*.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of Royal Statistics Society, Series B*, 39:1–38.
- Hojen-Sorensen, P., Winther, O., and Hansen, L. K. (2002). Mean-field approaches to independent component analysis. *Neural Computation*, 14:889–918.
- McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. John Wiley and Sons.
- Olsson, R. K., Lehn-Schiler, T., and Petersen, K. B. (2005). State-space models - from the em algorithm to a gradient approach. *Submitted to Neural Computation*.
- Petersen, K. B. and Winther, O. (2005a). Explaining slow convergence of EM in low noise linear mixtures. Technical Report 2005-2, Informatics and Mathematical Modelling, Technical University of Denmark.
- Petersen, K. B. and Winther, O. (2005b). The EM Algorithm in Independent Component Analysis. In *International Conference on Acoustics, Speech, and Signal Processing*.
- Salakhutdinov, R. and Roweis, S. (2003). Adaptive overrelaxed bound optimization methods. In *Proceedings of International Conference on Machine Learning, ICML*. International Conference on Machine Learning, ICML.

Salakhutdinov, R., Roweis, S., and Ghahramani, Z. (2003). Optimization with em and expectation-conjugate-gradient. In *Proceedings of International Conference on Machine Learning, ICML*. International Conference on Machine Learning, ICML.