On the Slow Convergence of EM and VBEM in Low-Noise Linear Models

Kaare Brandt Petersen
kbp@imm.dtu.dk
Ole Winther
owi@imm.dtu.dk
Lars Kai Hansen
lkhansen@imm.dtu.dk
Informatics and Mathematical Modeling, Technical University of Denmark, Building 321, DK = 2300 Kongens Lyngby, Denmark

We analyze convergence of the expectation maximization (EM) and variational Bayes EM (VBEM) 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, probabilistic principal component analysis, factor analysis, and Kalman filtering. Hence, the results are relevant for many practical applications.

1 Introduction

The expectation maximization (EM) algorithm introduced by Dempster, Laird, and Rubin (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 more accurate modeling of parameter uncertainty. EM convergence is known to slow dramatically when the signal-to-noise ratio is high, and a natural question is then: Will the more accurate modeling 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,

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

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 consist of the observed and unobserved data vectors as columns. The unobserved variables $\mathbf{S}$ are assumed to be distributed according to a prior $p(\mathbf{S})$, which can be gaussian (factor

analysis) or nongaussian (independent component analysis). The matrix $\mathbf{A}$ is referred to as the mixing matrix, and it can (but does not have to) be square ($m = d$). If $m < d$, we speak of the *overcomplete* case, while the opposite situation, $d < m$, is denoted *overdetermined*. In our discussion, the data are assumed prewhitened, that is, $\mathbf{X}\mathbf{X}^T = N\mathbf{I}$, a mild condition that simplifies the notation.

2 Slow Convergence in EM

For parameter estimation $(\mathbf{A}, \Sigma)$ 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 and then improve the parameters by maximizing the log likelihood averaged with regard to the approximate hidden variable posterior (for details, consult Dempster et al., 1977; McLachlan & Krishnan, 1997). Bermond and Cardoso (1999) made an important but seemingly little-known discovery about the convergence properties of the EM algorithm in the low-noise limit. 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 Figure 1) to obtain approximate expressions for the parameter updates. Using the notation $\mathbf{\Gamma} = \mathbf{X} - \mathbf{A}\mathbf{S}$, we get

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

where $\mathbf{A}_n$ denotes the estimated mixing matrix in the nth 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, $\sigma_{bias}^2 = 1 - \mathrm{rank}(\mathbf{A})/m$ and $z = \mathrm{rank}(\mathbf{A})/m - 2\mathrm{Tr}(\mathbf{U})/N$, where $\mathbf{U}$ is a data and prior dependent matrix. (This result is discussed in Petersen & Winther, 2005a, to which readers are 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 number of iterations are 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 & Sejnowski, 1995).

The slow convergence of the EM algorithm has been debated for a while, and many suggestions for speeding it up have been proposed (McLachlan & Krishnan, 1997). One straightforward method is to use a gradient-based

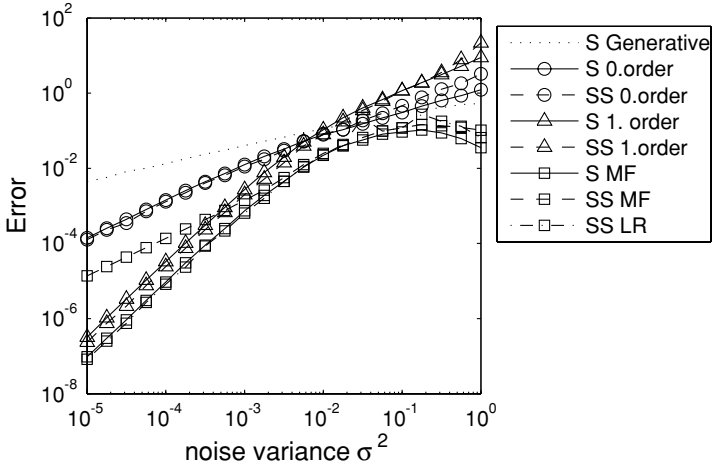

Figure 1: This plot demonstrates that the fundamental Taylor expansion of the moments $\langle \mathbf{S} \rangle$ and $\langle \mathbf{S}\mathbf{S}^T \rangle$ is reasonably accurate. A set of sources $\mathbf{S}_{gen}$ is generated using a mixture of gaussians (MoG) prior. From this, using a suitable 2×2 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 the approximation is fairly accurate when the noise variance is small. As expected, the first-order approximation (triangles) is more accurate than the zeroth-order approximation (circles) in the low-noise regime. 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 possible only for artificial data sets but is included for perspective. The mean field (MF) approximations to the posterior moments (squares) are also included for perspective (see Hojen-Sorensen, Winther, & Hansen, 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.

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, Lehn-Schiøler, & Petersen, 2005). Recently, another general technique, by Salakhudinov and Roweis (2003), called *adaptive overrelaxed EM*, was proposed, leading to considerable faster convergence (Petersen & Winther, 2005b). The key idea of the adaptive, overrelaxed EM is to boost the update by a factor $\eta \geq 1$. Combining this with the low-noise-limit analysis, we get

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

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

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, treating them at the same footing as the hidden variables. (See Beal & 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 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:

$$p(\mathbf{A}) \propto \exp\left[-\frac{1}{2}\mathrm{Tr}\left(\mathbf{A}\Sigma_p^{-1}\mathbf{A}^T\right)\right]$$
$$p(\sigma^2) \propto (\sigma^2)^{-(\alpha_p+1)} \exp\left[-\beta_p/\sigma^2\right].$$

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 $r^2 \to 0$, and we can expand the moments involved in updating the $\mathbf{A}$ pdf, in powers of r^2:

$$\langle \mathbf{A}\rangle_{n+1} = \mathbf{X}\langle \mathbf{S}\rangle^T\left[\langle \mathbf{S}\mathbf{S}^T\rangle + r^2\Sigma_p^{-1}\right]^{-1} = \langle \mathbf{A}\rangle_n + \mathcal{O}(r^2)$$
$$\mathrm{Var}(\mathbf{A})_{n+1} = r^2\left[\langle \mathbf{S}\mathbf{S}^T\rangle + r^2\Sigma_p^{-1}\right]^{-1} = \mathbf{0} + \mathcal{O}(r^2),$$

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

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

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, Rowels, and Ghahramani (2003), the convergence rate in a gaussian mixture model is demonstrated to be slow when the noise level is large, that is, when the mixtures have considerable overlap. The situation analyzed in this article, however, is a limit of low noise in which the problem intuitively should have an extraordinarily clear and well-defined solution. In that sense, this result is counterintuitive 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 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 suggest that this is not the core of the convergence problem and we are instead conjecturing that it is indeed the low-noise limit that is the essence of the matter.

Acknowledgments

The research for this note 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*. San Francisco: Morgan Kavffman.

Beal, M. J., & Ghahramani, Z. (2003). The variational Bayesian EM algorithm for incomplete data: With application to scoring graphical model structures. *Bayesian Statistics*, *7*, 453–465.

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

Bermond, O., & Cardoso, J. F. (1999). Approximate likelihood for noisy mixtures. In *Proceedings of the First International Workshop on Independent Component Analysis and Blind Source Separation, ICA '99*. Aussois, France.

Dempster, A. P., Laird, N. M., & 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., & Hansen, L. K. (2002). Mean-field approaches to independent component analysis. *Neural Computation, 14*, 889–918.

McLachlan, G. J., & Krishnan, T. (1997). *The EM algorithm and extensions*. New York: Wiley.

Olsson, R. K., Lehn-Schiøler, T., & Petersen, K. B. (2005). *State-space models—from the EM algorithm to a gradient approach*. Manuscript submitted for publication.

Petersen, K. B., & Winther, O. (2005a). *Explaining slow convergence of EM in low noise linear mixtures* (Tech. Rep. 2005-2). Kongens Lyngby: Informatics and Mathematical Modelling, Technical University of Denmark.

Petersen, K. B., & Winther, O. (2005b). The EM algorithm in independent component analysis. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Piscateway, NJ: IEEE.

Salakhutdinov, R., & Roweis, S. (2003). Adaptive overrelaxed bound optimization methods. In *Proceedings of International Conference on Machine Learning, ICML*. AAAI Press.

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

Received January 5, 2005; accepted March 2, 2005.