# The EM Algorithm in Independent Component Analysis

Kaare Brandt Petersen, Ole Winther
Technical University of Denmark, ISP Group

ICASSP, March 21, 2005

# Outline

- EM for ICA: Model, Cost and Optimization
- Numerical Example I & II
- Analysis of low noise limit (EM and VB-EM)
- Conclusions

# EM for ICA: Model and Cost

- **Model: Mean Field ICA** [Højen-Sørensen et al. 2002]
  Linear mixing and gaussian noise

$$\mathbf{x}_t = \mathbf{A}\mathbf{s}_t + \mathbf{n}_t$$
$$\mathbf{n}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}), \quad \mathbf{s}_t \sim p(\mathbf{s}_t|\boldsymbol{\nu}), \quad \boldsymbol{\theta} = \{\mathbf{A}, \mathbf{\Sigma}, \boldsymbol{\nu}\}$$

- **Cost function**
  Lower bound on log likelihood

$$B(\boldsymbol{\theta}, \phi) \leq \mathcal{L}(\boldsymbol{\theta}) = \ln p(\mathrm{Data}|\boldsymbol{\theta}), \qquad B(\boldsymbol{\theta}, \phi) = \int q(\mathbf{S}|\phi) \ln \frac{p(\mathbf{S}, \mathbf{X}|\boldsymbol{\theta})}{q(\mathbf{S}|\phi)} d\mathbf{S}$$

# EM for ICA: Optimization

- Expectation Maximization (**EM**)
  [Dempster et al., 1977]
  [Neal and Hinton, 1993]

$$\begin{aligned} &\textbf{E-step:} \quad \phi = \max_{\phi'} B(\boldsymbol{\theta}, \phi') \\ &\textbf{M-step:} \quad \boldsymbol{\theta} = \max_{\boldsymbol{\theta}'} B(\boldsymbol{\theta}', \phi) \end{aligned}$$

- Adaptive Overrelaxed EM (**Adap EM**)
  [Salakhutdinov and Roweis, 2003]

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

- **Easy Gradient Recipe**
  ”Gradients for the cost of EM steps”
  [Olsson et al., 2005]
  [Salakhutdinov et al., 2003]

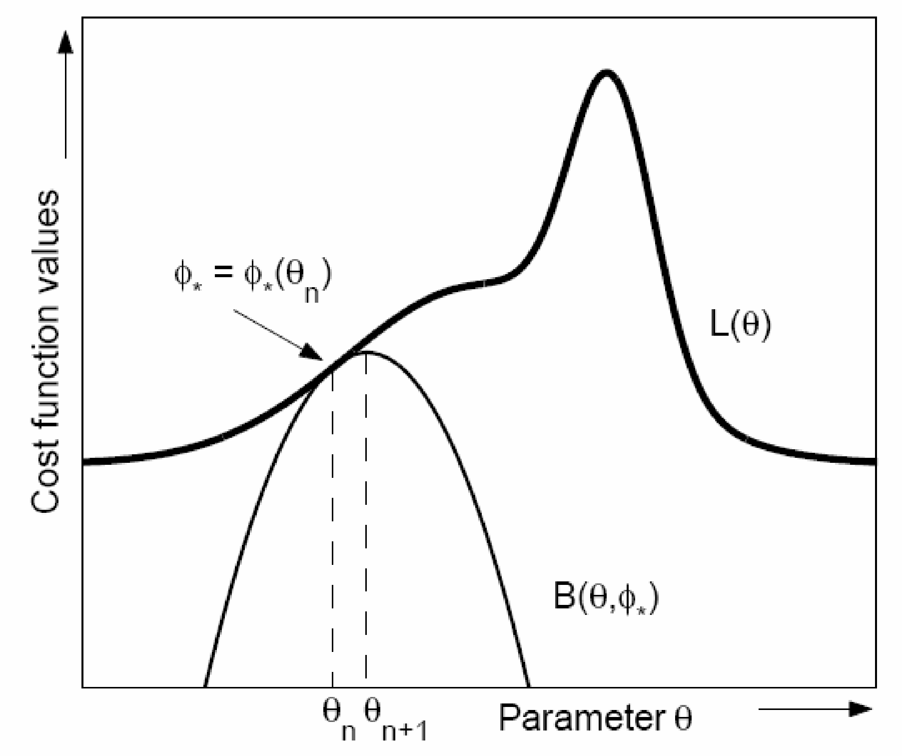
$$\begin{aligned} &\quad \texttt{function } [\mathcal{L}, \tfrac{d\mathcal{L}}{d\boldsymbol{\theta}}] = \texttt{loglikelihood}(\boldsymbol{\theta}) \\ &\textbf{1} \quad \text{Find } \phi^* \text{ such that } \tfrac{\partial B}{\partial \phi}\Big|_{\phi^*} = 0 \\ &\textbf{2} \quad \text{Calculate } \mathcal{L} = B(\boldsymbol{\theta}, \phi^*) \\ &\textbf{3} \quad \text{Calculate } \tfrac{d\mathcal{L}}{d\boldsymbol{\theta}} = \tfrac{\partial B}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}, \phi^*) \end{aligned}$$

Gradient based optimizers
e.g. UCMINF (BFGS w line search)

# EM for ICA: Bound and log likelihood

Important difference:

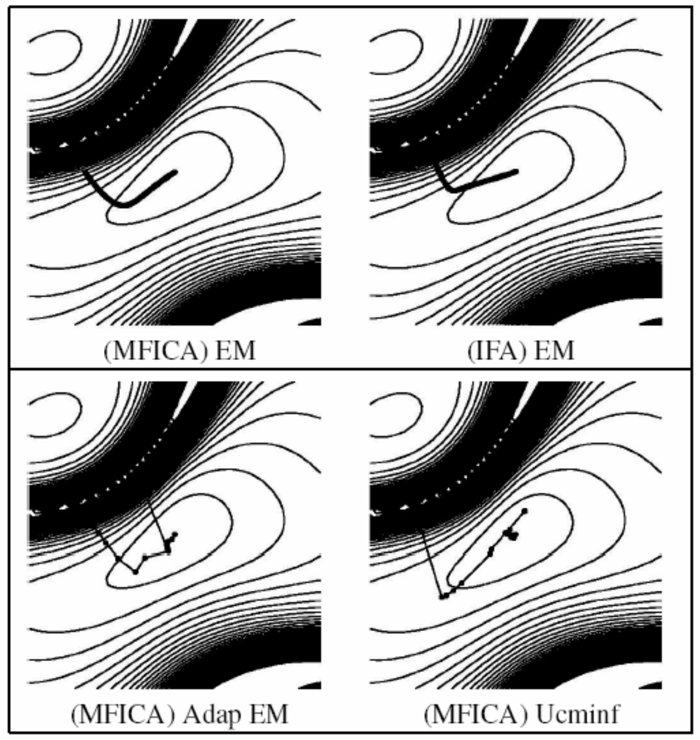
- EM is maximizing the bound
- Easy Gradient Recipe is maximizing the true log likelihood



# Numerical Example I

- Mean Field ICA
- Noise variance 0.01
- A 2x2 mixing matrix

| Opt. Method | Iterations |
|---|---:|
| EM | 729 |
| Adap EM | 16 |
| Easy Gradient | 25 |



# Numerical Example II

- Mean Field ICA
- Noise variance 0.01
- 300 random 2x2 mixing matrices

| **Opt. Method** | 0-50 Iterations | 50-100 Iterations | 100+ Iterations |
|---|---|---|---|
| EM | 0% | 2% | **98%** |
| Adap EM | 14% | **47%** | 39% |
| Easy Gradient | **93%** | 7% | 0% |

# Analysis: The Low Noise Limit

- **EM** for ICA [Bermond et al. 1999, Petersen and Winther 2005]
  Update of e.g. mixing matrix A

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

- **Variational Bayes EM** for ICA [Petersen et al. 2005]
  Update of hyperparameters (moments)
  of the mixing matrix distribution

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

where $r^2 = 1/\langle 1/s^2 \rangle$ is the order parameter

# Conclusions

- EM is very slow

  (Variational Bayes EM is also slow in the low noise limit)

- Nice alternatives:
  - **Adaptive Overrelaxed EM**
  - **Easy Gradient Recipe**

# Acknowledgements & References

**Thanks to**:

- Ole Winther, Lars Kai Hansen, Rasmus Olsson and Tue Lehn-Schiøler

**Relevant references:**

- K. B. Petersen, O. Winther, ”The EM Algorithm in Independent Component Analysis”, ICASSP 2005.
- K. B. Petersen, O. Winther, L. K. Hansen, ” On the slow convergence of EM and VBEM in low noise linear models”, *Neural Computation*, 2005
- K. B. Petersen and O. Winther, ”Explaining slow convergence of EM in low noise linear mixtures”, Tech-Rep-IMM 2-2005, 2005.
- R. K. Olsson, K. B. Petersen and T. Lehn-Schiøler, ”State-space models - from the EM algorithm to a gradient approach”, Submitted to Neural Computation, 2005.
- Salakhutdinov and Roweis, ”Adaptive Overrelaxed Bound Optimization Methods”, Proceedings of ICML, 2003.
- Salakhutdniov, Roweis and Gharamani, ”Optimization with EM and Expectation-Conjugate-Gradient”, Proceedings of ICML, 2003.