Statistical Models & Computing Methods

Lecture 11: Advanced EM

Cheng Zhang

School of Mathematical Sciences, Peking University

November 4, 2024

Introduction

- While EM increases the marginal likelihood in each iteration and often converges to a stationary point, we are not clear about the convergence rate and how does that relate to the missing data scenario.
- Moreover, the requirements of tractable conditional distribution and easy complete data MLE may be too restrictive in practice.
- In this lecture, we will discuss the convergence theory for EM and introduce some variants of it that can be applied in more general settings.

Example: Censored Survival Times

- Recall that in the censored survival times example, given the observed data $Y = \{(t_1, \delta_1), \ldots, (t_n, \delta_n)\}$, where t_j follows an exponential distribution with mean μ and can be either censored or not as indicated by δ_j.
- Assume $\delta_i = 0, i \leq r,\ \delta_i = 1, i > r$. The MLE of μ is $\hat{\mu} = \sum_{i=1}^n t_i / r$
- EM update formula

$$\mu^{(k+1)} = \frac{\sum_{i=1}^n t_i + (n-r)\mu^{(k)}}{n}$$

- Therefore,

$$\mu^{(k+1)} - \hat{\mu} = \frac{n-r}{n}(\mu^{(k)} - \hat{\mu})$$

  Linear convergence, rate depends on the amount of missing information

EM as A Fixed Point Algorithm

We can view EM update as a map

$$\theta^{(t+1)} = \Phi(\theta^{(t)}), \quad \Phi(\theta) = \arg\max_{\theta'} Q(\theta'|\theta)$$

where $Q(\theta'|\theta) = \mathbb{E}_{p(z|x,\theta)} \log p(x, z|\theta')$

Lemma 1

If for some θ^*, $\mathcal{L}(\theta^*) \geq \mathcal{L}(\theta)$, $\forall\theta$, then for every EM algorithm

$$\mathcal{L}(\Phi(\theta^*)) = \mathcal{L}(\theta^*), \; Q(\Phi(\theta^*)|\theta^*) = Q(\theta^*|\theta^*)$$

and

$$p(z|x, \Phi(\theta^*)) = p(z|x, \theta^*), \text{ a.s.}$$

Lemma 2
If for some θ^*, $\mathcal{L}(\theta^*) > \mathcal{L}(\theta)$, $\forall \theta \neq \theta^*$, then for every EM algorithm

$$\Phi(\theta^*) = \theta^*$$

Theorem 1
Suppose that $\theta^{(t)}, t = 0, 1, \ldots$ is an instance of an EM algorithm such that

- the sequence $\mathcal{L}(\theta^{(t)})$ is bounded and $\mathcal{L}(\theta^{(t)}) \to \mathcal{L}^*$.
- $\mathcal{M}(\mathcal{L}^*) = \{\theta^* : \mathcal{L}(\theta^*) = \mathcal{L}^*\}$ is a discrete set.
- $\|\theta^{(t+1)} - \theta^{(t)}\| \to 0$.

Then all the limit points of the sequence $\theta^{(t)}$ converges to some $\theta^* \in \mathcal{M}(\mathcal{L}^*)$. See Wu (1983) for more details.

Local Convergence

- Since $\theta^{(t+1)} = \Phi(\theta^{(t)})$ maximizes $Q(\theta'|\theta^{(t)})$, we have

$$\frac{\partial Q}{\partial \theta'}(\theta^{(t+1)}|\theta^{(t)}) = 0$$

- For all t, there exists a $0 \leq \alpha_0^{(t+1)} \leq 1$ such that

$$Q(\theta^{(t+1)}|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}) = -(\theta^{(t+1)} - \theta^{(t)}) \cdot \frac{\partial^2 Q}{\partial \theta'^2}(\theta_0^{(t+1)}|\theta^{(t)})(\theta^{(t+1)} - \theta^{(t)})^T$$

  where $\theta_0^{(t+1)} = \alpha_0 \theta^{(t)} + (1 - \alpha_0)\theta^{(t+1)}$

- If the sequence $\frac{\partial^2 Q}{\partial \theta'^2}(\theta_0^{(t+1)}|\theta^{(t)})$ is negative definite with eigenvalues bounded away from zero and $\mathcal{L}(\theta^{(t)})$ is bounded, by Theorem 1, $\theta^{(t)}$ converges to some θ^*

Local Convergence

- When EM converges, it converges to a fixed point of the map

$$\theta^* = \Phi(\theta^*)$$

- Taylor expansion of Φ at θ^* yields

$$\theta^{(t+1)} - \theta^* = \Phi(\theta^{(t)}) - \Phi(\theta^*) \approx \nabla\Phi(\theta^*)(\theta^{(t)} - \theta^*)$$

- The global rate of EM defined as

$$\rho = \lim_{t\to\infty} \frac{\|\theta^{(t+1)} - \theta^*\|}{\|\theta^{(t)} - \theta^*\|}$$

  equals the largest eigenvalue of $\nabla\Phi(\theta^*)$ and $\rho < 1$ when the observed Fisher information $-\nabla^2\mathcal{L}(\theta^*)$ is positive definite.

Proof

- As aforementioned, $\Phi(\theta)$ maximize $Q(\theta'|\theta)$, therefore

$$\frac{\partial Q}{\partial \theta'}(\Phi(\theta)|\theta) = 0, \quad \forall \theta$$

- Differentiate w.r.t. θ

$$\frac{\partial^2 Q}{\partial \theta'^2}(\Phi(\theta)|\theta)\nabla\Phi(\theta) + \frac{\partial^2 Q}{\partial \theta \partial \theta'}(\Phi(\theta)|\theta) = 0$$

let $\theta = \theta^*$

$$\nabla\Phi(\theta^*) = \left(-\frac{\partial^2 Q}{\partial \theta'^2}(\theta^*|\theta^*)\right)^{-1} \frac{\partial^2 Q}{\partial \theta \partial \theta'}(\theta^*|\theta^*) \tag{1}$$

- If $\frac{\partial^2 Q}{\partial \theta'^2}(\theta^{(t+1)}|\theta^{(t)})$ is negative definite with eigenvalues bounded away from zero, then

$$-\frac{\partial^2 Q}{\partial \theta'^2}(\theta^*|\theta^*) = \mathbb{E}_{p(z|x,\theta^*)}\left(-\nabla^2 \log p(x,z|\theta^*)\right)$$

is positive definite, known as the **complete information**
- The marginal log-likelihood can be rewritten as

$$\begin{aligned}\mathcal{L}(\theta') &= \mathbb{E}_{p(z|x,\theta)} \log p(x,z|\theta') - \mathbb{E}_{p(z|x,\theta)} \log p(z|x,\theta') \\ &= Q(\theta'|\theta) - H(\theta'|\theta)\end{aligned}$$

Therefore

$$\frac{\partial^2 Q}{\partial \theta \partial \theta'}(\theta'|\theta) = \frac{\partial^2 H}{\partial \theta \partial \theta'}(\theta'|\theta)$$

- Some properties of $H(\theta|\theta) = \mathbb{E}_{p(z|x,\theta)} \log p(z|x,\theta)$

$$\frac{\partial H}{\partial \theta'}(\theta|\theta) = 0$$
$$\frac{\partial^2 H}{\partial \theta \partial \theta'}(\theta|\theta) = -\frac{\partial^2 H}{\partial \theta'^2}(\theta|\theta)$$

- Therefore,

$$\frac{\partial^2 Q}{\partial \theta \partial \theta'}(\theta^*|\theta^*) = \frac{\partial^2 H}{\partial \theta \partial \theta'}(\theta^*|\theta^*) = -\frac{\partial^2 H}{\partial \theta'^2}(\theta^*|\theta^*)$$

is positive semidefinite (variance of the score $\nabla \log p(z|x,\theta^*)$), known as the **missing information**

Missing-Information Principle

$$\mathcal{L}(\theta') = Q(\theta'|\theta) - H(\theta'|\theta)$$

- Differentiate both side w.r.t. θ' twice

$$\nabla^2 \mathcal{L}(\theta') = \frac{\partial^2 Q}{\partial \theta'^2}(\theta'|\theta) - \frac{\partial^2 H}{\partial \theta'^2}(\theta'|\theta)$$

- The *missing-information principle*

$$\underbrace{-\frac{\partial^2 Q}{\partial \theta'^2}(\theta|\theta)}_{I_{\text{complete}}} = \underbrace{-\nabla^2 \mathcal{L}(\theta)}_{I_{\text{observed}}} + \underbrace{-\frac{\partial^2 H}{\partial \theta'^2}(\theta|\theta)}_{I_{\text{missing}}}$$

- Substitute in (1)

$$\begin{aligned}\nabla \Phi(\theta^*) &= I_{\text{complete}}^{-1}(\theta^*) I_{\text{missing}}(\theta^*) \\ &= \left(I_{\text{observed}}(\theta^*) + I_{\text{missing}}(\theta^*)\right)^{-1} I_{\text{missing}}(\theta^*)\end{aligned}$$

Convergence Rate of EM

- When $I_{\text{observed}} = -\nabla^2 \mathcal{L}(\theta^*)$ is positive definite, the eigenvalues of $\nabla \Phi(\theta^*)$ are all less than 1, EM has a linear convergence rate.
- The rate of convergence depends on the relative size of $I_{\text{observed}}(\theta^*)$ and $I_{\text{missing}}(\theta^*)$. EM converges rapidly when the missing information is small.
- The fraction of information loss may vary across different component of θ, so some component may converge faster than other components.
- See Wu (1983) for more detailed discussions.

EM for Maximum A Posterior

- EM can be easily modified for the Maximum A Posterior (MAP) estimate instead of the MLE.
- Suppose the log-prior penalty term is $R(\theta)$. We only have to maximize

$$Q(\theta|\theta^{(t)}) + R(\theta) \tag{2}$$

  in the M-step
- Monotonicity.

$$\begin{aligned}
\mathcal{L}(\theta^{(t+1)}) + R(\theta^{(t+1)}) &\geq \mathcal{F}(\theta^{(t+1)}|\theta^{(t)}) + R(\theta^{(t+1)}) \\
&\geq \mathcal{F}(\theta^{(t)}|\theta^{(t)}) + R(\theta^{(t)}) \\
&= \mathcal{L}(\theta^{(t)}) + R(\theta^{(t)})
\end{aligned}$$

- If $R(\theta)$ corresponds to a conjugate prior, (2) can be maximized in the same manner as $Q(\theta|\theta^{(t)})$.

- The E-step requires finding the expected complete data log-likelihood $Q(\theta|\theta^{(t)})$. When this expectation is difficult to compute, we can approximate it via Monte Carlo methods
- **Monte Carlo EM** (Wei and Tanner, 1990)
  - Draw missing data $z_1^{(t)}, \ldots, z_m^{(t)}$ from the conditional distribution $p(z|x, \theta^{(t)})$
  - Compute a Monte Carlo estimate of $Q(\theta|\theta^{(t)})$

$$\hat{Q}^{(t+1)}(\theta|\theta^{(t)}) = \frac{1}{m} \sum_{i=1}^{m} \log p(x, z_i^{(t)}|\theta)$$

  - Update $\theta^{(t+1)}$ to maximize $\hat{Q}^{(t+1)}(\theta|\theta^{(t)})$.

Remark: It is recommended to let m changes along iterations (small at the beginning and increases as iterations progress)

- By the lack of memory, it is easy to compute the expected complete data log-likelihood, which lead to the ordinary EM update

$$\mu_{\mathrm{EM}}^{(k+1)} = \frac{\sum_{i=1}^{n} t_i + (n-r)\mu^{(k)}}{n}$$

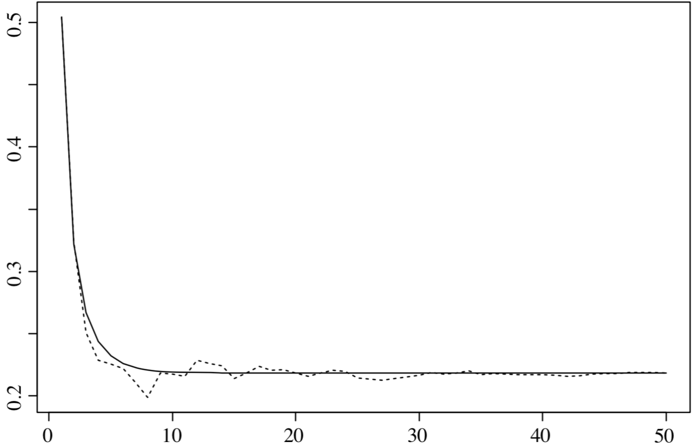
- In MCEM, we can sample from the conditional distribution

$$\boldsymbol{T}_j = (T_{j,r+1}, \ldots, T_{j,n}),\ T_{j,l} - t_l \sim \mathrm{Exp}(\mu^{(k)}), \quad l = r+1, \ldots, n$$

for $j = 1, \ldots, m^{(k)}$, and the update formula is

$$\mu_{\mathrm{MCEM}}^{(k+1)} = \frac{\sum_{i=1}^{n} t_i + \frac{1}{m^{(k)}} \sum_{j=1}^{m^{(k)}} \boldsymbol{T}_j^T \mathbf{1}}{n}$$

Examples: Censored Survival Times



Improving the M-step

- One of the appeals of the EM algorithm is that $Q(\theta|\theta^{(t)})$ is often simpler to maximize than the marginal likelihood
- In some cases, however, the M-step cannot be carried out easily even though the computation of $Q(\theta|\theta^{(t)})$ is straightforward in the E-step
- For such situations, Dempster et al (1977) defined a generalized EM algorithm (GEM) for which the M-step only requires $\theta^{(t+1)}$ to improve $Q(\theta|\theta^{(t)})$

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta^{(t+1)}|\theta^{(t)})$$

- We can easily show that GEM is also monotonic in $\mathcal{L}$

$$\mathcal{L}(\theta^{(t+1)}) \geq \mathcal{F}(q^{(t)}, \theta^{(t+1)}) \geq \mathcal{F}(q^{(t)}, \theta^{(t)}) = \mathcal{L}(\theta^{(t)})$$

- Meng and Rubin (1993) replaces the M-step with a series of computationally cheaper conditional maximization (CM) steps, leading to the **ECM** algorithm
- The M-step in ECM contains a collection of simple CM steps, called a CM *cycle*. For $s = 1, \ldots, S$, the s-th CM step requires the maximization of $Q(\theta|\theta^{(t)})$ subject to a constraint

$$\theta^{(t+s/S)} = \arg\max_{\theta} Q(\theta|\theta^{(t)}), \quad \text{s.t. } g_s(\theta) = g_s(\theta^{(t+(s-1)/S)})$$

- The efficiency of ECM depends on the choice of constraints. Examples: Blockwise updates (coordinate ascent).
- One may also insert an E-step between each pair of CM-steps, updating Q at every stage of the CM cycle.

Multivariate Regression

- Suppose we have n independent observations from the following k-variate normal model

$$Y_i \sim \mathcal{N}(X_i\beta, \Sigma), \quad i = 1, \ldots, n$$

  - $X_i \in \mathbb{R}^{k \times p}$ is a known design matrix for the i-th observation
  - β is a vector of p unknown parameters
  - Σ is a $d \times d$ unknown variance-covariance matrix
- The complete data log-likelihood (up to a constant) is

$$L(\beta, \Sigma|Y) = -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(Y_i - X_i\beta)^T\Sigma^{-1}(Y_i - X_i\beta)$$

- Generally, MLE does not has closed form solution except in special cases (e.g., $\Sigma = \sigma^2 I$)

A Coordinate Ascent Algorithm

- Although the joint maximization of β and Σ are not generally in closed form, a coordinate ascent algorithm does exist
- Given $\Sigma = \Sigma^{(t)}$, the conditional MLE of β is simply the weighted least-square estimate

$$\beta^{(t+1)} = \left(\sum_{i=1}^{n} X_i^T (\Sigma^{(t)})^{-1} X_i\right)^{-1} \left(\sum_{i=1}^{n} X_i^T (\Sigma^{(t)})^{-1} Y_i\right)$$

- Given $\beta = \beta^{(t+1)}$, the conditional MLE of Σ is the cross-product of the residuals

$$\Sigma^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i\beta^{(t+1)})(Y_i - X_i\beta^{(t+1)})^T$$

- Now suppose that we also have missing data

$$Y_i \sim \mathcal{N}(X_i\beta, \Sigma), \quad i = n+1, \ldots, m$$

for which only the design matrix $X_i,\ i > n$ are known

- The complete data log-likelihood

$$L(\beta, \Sigma | Y) = -\frac{m}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{m}(Y_i - X_i\beta)^T\Sigma^{-1}(Y_i - X_i\beta)$$

- Expected values of sufficient statistics observed data and current parameter $\theta^{(t)} = (\beta^{(t)}, \Sigma^{(t)})$

$$\mathbb{E}(Y_i | Y_{\mathrm{obs}}, \theta^{(t)}) = X_i\beta^{(t)}$$
$$\mathbb{E}(Y_iY_i^T | Y_{\mathrm{obs}}, \theta^{(t)}) = \Sigma^{(t)} + (X_i\beta^{(t)})(X_i\beta^{(t)})^T$$

Expected complete-data log-likelihood

$$
\begin{aligned}
Q(\theta|\theta^{(t)}) &= -\frac{m}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(Y_i - X_i\beta)^T\Sigma^{-1}(Y_i - X_i\beta) \\
&\quad - \frac{1}{2}\sum_{i=n+1}^{m}\mathbb{E}\left((Y_i - X_i\beta)^T\Sigma^{-1}(Y_i - X_i\beta)\right) \\
&= -\frac{m}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(Y_i - X_i\beta)^T\Sigma^{-1}(Y_i - X_i\beta) \\
&\quad - \frac{1}{2}\sum_{i=n+1}^{m}(\mathbb{E}Y_i - X_i\beta)^T\Sigma^{-1}(\mathbb{E}Y_i - X_i\beta) + C
\end{aligned}
$$

where $C = \frac{1}{2}\sum_{i=n+1}^{m}\mathbb{E}(Y_i)^T\Sigma^{-1}\mathbb{E}(Y_i) - \mathbb{E}(Y_i^T\Sigma^{-1}Y_i)$ is a constant independent of the parameter β.

CM-steps: Update β

- The first CM-step, maximize Q given $\Sigma = \Sigma^{(t)}$.
- Since C is independent of β, we can maximize

$$\begin{aligned} -\frac{m}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(Y_i - X_i\beta)^T\Sigma^{-1}(Y_i - X_i\beta) \\ -\frac{1}{2}\sum_{i=n+1}^{m}(\mathbb{E}Y_i - X_i\beta)^T\Sigma^{-1}(\mathbb{E}Y_i - X_i\beta) \end{aligned}$$

$$\Rightarrow \beta^{(t+1)} = \left(\sum_{i=1}^{m} X_i^T\Sigma^{(t)}X_i\right)^{-1}\left(\sum_{i=1}^{m} X_i^T\Sigma^{(t)}\hat{Y}_i\right)$$

where

$$\hat{Y}_i = \begin{cases} Y_i, & i \le n \\ X_i\beta^{(t)}, & i > n \end{cases}$$

CM-steps: Update Σ

- The second CM-step, maximize Q with $\beta = \beta^{(t+1)}$
- Rewrite Q as

$$Q(\theta|\theta^{(t)}) = \frac{m}{2}\log|\Sigma^{-1}| - \frac{1}{2}\sum_{i=1}^{n}\mathrm{Tr}\left(\Sigma^{-1}(Y_i - X_i\beta)(Y_i - X_i\beta)^T\right)$$
$$- \frac{1}{2}\sum_{i=n+1}^{m}\mathrm{Tr}\left(\Sigma^{-1}\mathbb{E}\left((Y_i - X_i\beta)(Y_i - X_i\beta)^T\right)\right)$$

- Similarly as in the complete data case

$$\Sigma^{(t+1)} = \frac{1}{m}\left(\sum_{i=1}^{n}(Y_i - X_i\beta^{(t+1)})(Y_i - X_i\beta^{(t+1)})^T + \sum_{i=n+1}^{m}\Sigma^{(t)}\right.$$
$$\left. + \sum_{i=n+1}^{m} X_i(\beta^{(t)} - \beta^{(t+1)})(\beta^{(t)} - \beta^{(t+1)})^T X_i^T\right)$$

- Both the E-step and the two CM-steps can be implemented using close form solutions, no numerical iteration required.
- Both CM-steps improves Q

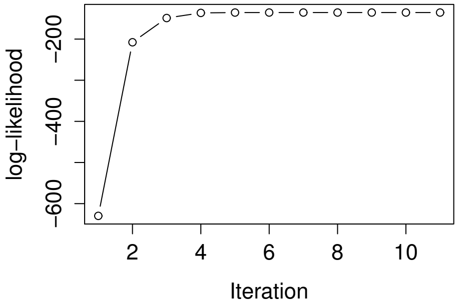
$$
\begin{aligned}
Q(\beta^{(t+1)}, \Sigma^{(t+1)} | \beta^{(t)}, \Sigma^{(t)}) &\geq Q(\beta^{(t+1)}, \Sigma^{(t)} | \beta^{(t)}, \Sigma^{(t)}) \\
&\geq Q(\beta^{(t)}, \Sigma^{(t)} | \beta^{(t)}, \Sigma^{(t)})
\end{aligned}
$$

- ECM in this case can be viewed as an efficient generalization of iterative reweighted least squares, in the presence of missing data.

We generate 120 design matrices at random and simulate 100 observations with $\beta = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 1, & 0.1 \\ 0.1 & 2 \end{pmatrix}$

ECM estimates

$$\hat{\beta} = \begin{pmatrix} 2.068 \\ 1.087 \end{pmatrix}, \quad \hat{\Sigma} = \begin{pmatrix} 0.951 & 0.214 \\ 0.214 & 2.186 \end{pmatrix}$$



EM Gradient Algorithm

- Iterative optimization can be considered when direct maximization is not available.
- All numerical optimization can apply and that would yield an algorithm that has nested iterative loops (e.g., ECM inserts conditional maximization steps within each CM cycle)
- To avoid the computational burden of nested looping, Lange proposed to use one single step of Newton's method

$$\begin{aligned}\theta^{(t+1)} &= \theta^{(t)} - \left(\frac{\partial^2 Q}{\partial \theta'^2}(\theta^{(t)}|\theta^{(t)})\right)^{-1} \frac{\partial Q}{\partial \theta'}(\theta^{(t)}|\theta^{(t)}) \\ &= \theta^{(t)} - \left(\frac{\partial^2 Q}{\partial \theta'^2}(\theta^{(t)}|\theta^{(t)})\right)^{-1} \nabla \mathcal{L}(\theta^{(t)})\end{aligned}$$

- This EM gradient algorithm has the same rate of convergence as the full EM algorithm.

- When EM is slow, we can use the relatively simple analytic setup from EM to motivate particular forms for Newton-like steps.
- **Aitken Acceleration**. Newton update

$$\theta^{(t+1)} = \theta^{(t)} - (\nabla^2 \mathcal{L}(\theta^{(t)}))^{-1} \nabla \mathcal{L}(\theta^{(t)}) \tag{3}$$

  Note that $\nabla \mathcal{L}(\theta^{(t)}) = \frac{\partial Q}{\partial \theta'}(\theta^{(t)}|\theta^{(t)})$ and

$$0 = \frac{\partial Q}{\partial \theta'}(\theta_{\text{EM}}^{(t+1)}|\theta^{(t)}) \approx \frac{\partial Q}{\partial \theta'}(\theta^{(t)}|\theta^{(t)}) + \frac{\partial^2 Q}{\partial \theta'^2}(\theta^{(t)}|\theta^{(t)})(\theta_{\text{EM}}^{(t+1)} - \theta^{(t)})$$

  substitute in (3)

$$\theta^{(t+1)} = \theta^{(t)} + (I_{\text{observed}}(\theta^{(t)}))^{-1} I_{\text{complete}}(\theta^{(t)})(\theta_{\text{EM}}^{(t+1)} - \theta^{(t)})$$

- Many other acceleration exists (e.g., Quasi-Newton methods).

- C. F. J. Wu. On the convergence properties of the EM algorithm. Annals of Statistics, 11:95–103, 1983.
- X.-L. Meng and D. B. Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika, 80:267–278, 1993.
- G. C. G. Wei and M. A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. Journal of the American Statistical Association, 85:699–704, 1990.
- K. Lange. A gradient algorithm locally equivalent to the EM algorithm. Journal of the Royal Statistical Society, Series B, 57:425–437, 1995.

References

- T. A. Louis. Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society, Series B, 44:226–233, 1982.