Computational statistics

Lecture 5: EM algorithm

Thierry Denœux

24 March, 2016

EM Algorithm

- An iterative optimization strategy motivated by a notion of missingness and by consideration of the conditional distribution of what is missing given what is observed.
- Can be very simple to implement. Can reliably find an optimum through stable, uphill steps.
- Difficult likelihoods often arise when data are missing. EM simplifies such problems. In fact, the 'missing data' may not truly be missing: they may be only a conceptual ploy to exploit the EM simplification!

Overview

Notation

$\mathbf{Y}$: Observed variables.

$\mathbf{Z}$: Missing or latent variables.

$\mathbf{X}$: Complete data $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$.

$\boldsymbol{\theta}$: Unknown parameter.

$L(\boldsymbol{\theta})$: observed-data likelihood, short for $L(\boldsymbol{\theta}; \mathbf{y}) = f(\mathbf{y}; \boldsymbol{\theta})$

$L_c(\boldsymbol{\theta})$: complete-data likelihood, short for $L(\boldsymbol{\theta}; \mathbf{x}) = f(\mathbf{x}; \boldsymbol{\theta})$

$\ell(\boldsymbol{\theta}), \ell_c(\boldsymbol{\theta})$: observed and complete-data log-likelihoods.

Notation

- Suppose we seek to maximize $L(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.
- Define $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ to be the expectation of the complete-data log-likelihood, conditional on the observed data $\mathbf{Y} = \mathbf{y}$. Namely

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) =& \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left\{ \ell_c(\boldsymbol{\theta}) \mid \mathbf{y} \right\} \\ =& \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left\{ \log f_{\mathbf{X}}(\mathbf{X}; \boldsymbol{\theta}) \mid \mathbf{y}, \boldsymbol{\theta}^{(t)} \right\} \\ =& \int \left[\log f_{\mathbf{X}}(\mathbf{x}) \right] f_{\mathbf{Z}|\mathbf{Y}}(\mathbf{z}|\mathbf{y}) \, d\mathbf{z} \end{aligned}$$

where the last equation emphasizes that $\mathbf{Z}$ is the only random part of $\mathbf{X}$ once we are given $\mathbf{Y} = \mathbf{y}$.

The EM Algorithm

Start with $\boldsymbol{\theta}^{(0)}$. Then

1. **E step**: Compute $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$.
2. **M step**: Maximize $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$ with respect to $\boldsymbol{\theta}$. Set $\boldsymbol{\theta}^{(t+1)}$ equal to the maximizer of Q.
3. Return to the E step unless a stopping criterion has been met; e.g.,

$$L(\boldsymbol{\theta}^{(t+1)}) - L(\boldsymbol{\theta}^{(t)}) \leq \epsilon$$

Convergence of the EM Algorithm

- It can be proved that $L(\boldsymbol{\theta})$ increases after each EM iteration, i.e., $L(\boldsymbol{\theta}^{(t+1)}) \geq L(\boldsymbol{\theta}^{(t)})$ for $t = 0, 1, \ldots$.
- Consequently, the algorithm converges to a local maximum of $L(\boldsymbol{\theta})$ if the likelihood function is bounded above.

Trival example

- Y, Z iid from $\mathcal{E}(\theta)$ with $y = 5$ observed but z missing.
- The complete-data log likelihood function is

$$\ell_c(\theta) = \log\{f_{\mathbf{X}}(\mathbf{x}; \theta)\} = 2\log(\theta) - \theta y - \theta z.$$

- Thus

$$Q(\theta, \theta^{(t)}) = 2\log(\theta) - 5\theta - \theta/\theta^{(t)}$$

since $\mathbb{E}_{\theta^{(t)}}\{Z|y\} = \mathbb{E}_{\theta^{(t)}}\{Z\} = 1/\theta^{(t)}$ follows from independence.

- The maximizer of $Q(\theta, \theta^{(t)})$ is the root of $2/\theta - 5 - 1/\theta^{(t)} = 0$. Thus $\theta^{(t+1)} = \frac{2\theta^{(t)}}{5\theta^{(t)}+1}$. Converges quickly to $\hat{\theta} = 0.2$.
- This example is not realistic. Easy analytic solution. Taking the required expectation is tricker in real applications because one needs to know the conditional distribution of the complete data given the missing data.

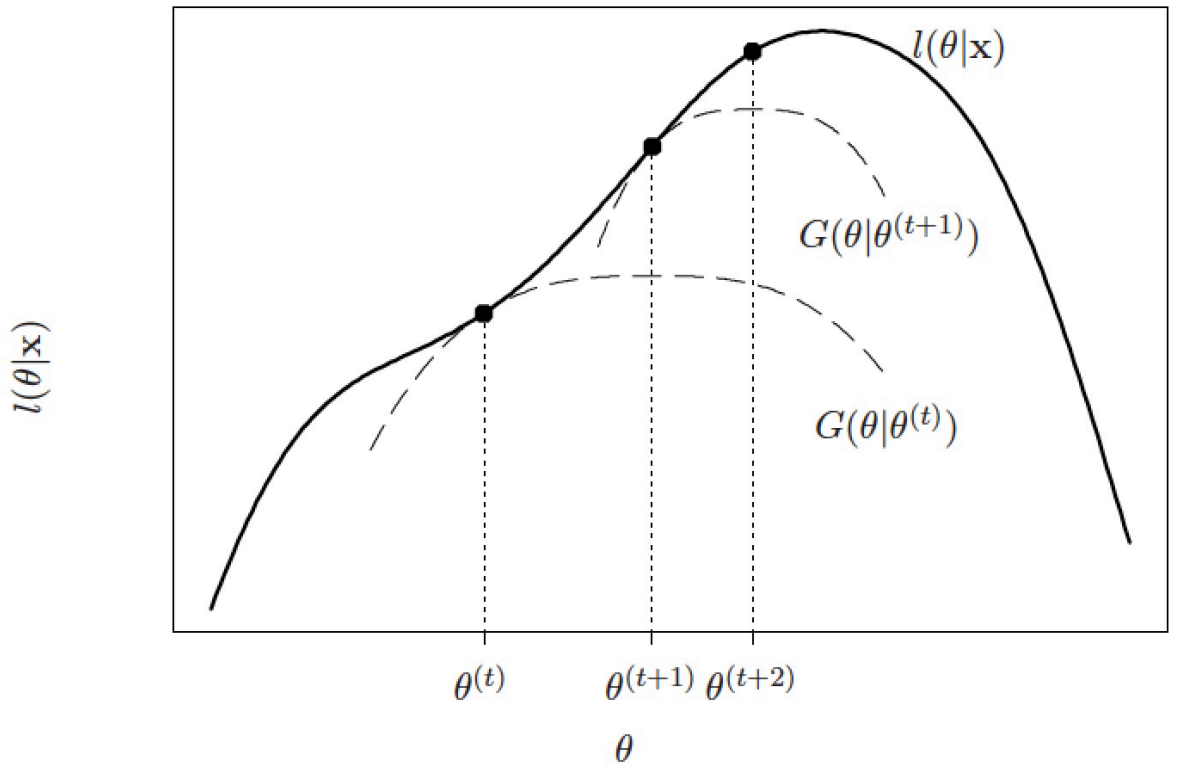
The nature of EM

- Ascent: Each M-step increases the log likelihood.
- Convergence: is linear (slow!). Rate is inversely related to the proportion of missing data.
- Optimization transfer:

$$\ell(\boldsymbol{\theta}|\mathbf{y}) \geq Q(\boldsymbol{\theta}, \theta^{(t)}) + \ell(\boldsymbol{\theta}^{(t)}|\mathbf{y}) - Q(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t)}) = G(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}).$$

- The last two terms in $G(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$ are constant with respect to $\boldsymbol{\theta}$, so Q and G are maximized at the same $\boldsymbol{\theta}$.
- Further, G is tangent to ℓ at $\boldsymbol{\theta}^{(t)}$, and lies everywhere below ℓ. We say that G is a minorizing function for l.
- EM transfers optimization from ℓ to the surrogate function G, which is more convenient to maximize.

The nature of EM



One-dimensional illustration of EM algorithm as a minorization or optimization transfer strategy. Each E step forms a minorizing function G, and each M step maximizes it to provide an uphill step.

Bayesian posterior mode

- Consider a Bayesian estimation problem with likelihood $L(\theta)$ and priori $f(\theta)$.
- The posterior density if proportional to $L(\theta)f(\theta)$. It can also be maximized by the EM algorithm.
- The E-step requires

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left\{ \ell_c(\boldsymbol{\theta}) \mid \mathbf{y} \right\} + \log f(\theta)$$

- The addition of the log-prior often makes it more difficult to maximize Q during the M-step.
- Some methods can be used to facilitate the M-step in difficult situations (see below).

Overview

Monte Carlo EM (MCEM)

- Replace the tth E step with
  1. Draw missing datasets $\mathbf{Z}_1^{(t)}, \ldots, \mathbf{Z}_{m^{(t)}}^{(t)}$ i.i.d. from $f_{\mathbf{Z}|\mathbf{Y}}(\mathbf{z}|\mathbf{y}; \boldsymbol{\theta}^{(t)})$. Each $\mathbf{Z}_j^{(t)}$ is a vector of all the missing values needed to complete the observed dataset, so $\mathbf{X}_j^{(t)} = (\mathbf{y}, \mathbf{Z}_j^{(t)})$ denotes a completed dataset where the missing values have been replaced by $\mathbf{Z}_j^{(t)}$.
  2. Calculate $\hat{Q}^{(t+1)}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = \frac{1}{m^{(t)}} \sum_{j=1}^{m^{(t)}} \log f_{\mathbf{X}}(\mathbf{X}_j^{(t)}|\boldsymbol{\theta})$.
- Then $\hat{Q}^{(t+1)}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$ is a Monte Carlo estimate of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$.
- The M step is modified to maximize $\hat{Q}^{(t+1)}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$.
- Increase $m^{(t)}$ as iterations progress to reduce the Monte Carlo variability of $\hat{Q}$. MCEM will not converge in the same sense as ordinary EM, rather values of $\boldsymbol{\theta}^{(t)}$ will bounce around the true maximum, with a precision that depends on $m^{(t)}$.

ECM algorithm

- Replaces the M step with a series of computationally simpler conditional maximization (CM) steps.
- Call the collection of simpler CM steps after the tth E step a CM cycle. Thus, the tth iteration of ECM is comprised of the tth E step and the tth CM cycle. r
- Let S denote the total number of CM steps in each CM cycle.

ECM algorithm (continued)

- For $s = 1, \ldots, S$, the sth CM step in the tth cycle requires the maximization of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$ subject to (or conditional on) a constraint, say

$$\mathbf{g}_s(\boldsymbol{\theta}) = \mathbf{g}_s(\boldsymbol{\theta}^{(t+(s-1)/S)})$$

  where $\boldsymbol{\theta}^{(t+(s-1)/S)}$ is the maximizer found in the $(s-1)$th CM step of the current cycle.
- When the entire cycle of S steps of CM has been completed, we set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t+S/S)}$ and proceed to the E step for the $(t+1)$th iteration.
- The art of constructing an effective ECM algorithm lies in choosing the constraints cleverly.

Choice 1: Iterated Conditional Modes / Gauss-Seidel

- Partition $\boldsymbol{\theta}$ into S subvectors, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_S)$.
- In the sth CM step, maximize Q with respect to $\boldsymbol{\theta}_s$ while holding all other components of $\boldsymbol{\theta}$ fixed.
- This amounts to the constraint induced by the function

$$g_s(\boldsymbol{\theta}) = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{s-1}, \boldsymbol{\theta}_{s+1}, \ldots, \boldsymbol{\theta}_S).$$

Choice 2

- At the sth CM step, maximize Q with respect to all other components of $\boldsymbol{\theta}$ while holding $\boldsymbol{\theta}_s$ fixed.
- Then $g_s(\boldsymbol{\theta}) = \boldsymbol{\theta}_s$.
- Additional systems of constraints can be imagined, depending on the particular problem context.
- A variant of ECM inserts an E step between each pair of CM steps, thereby updating Q at every stage of the CM cycle.

EM gradient algorithm

- Replace the M step with a single step of Newton's method, thereby approximating the maximum without actually solving for it exactly.
- Instead of maximizing, choose:

$$\begin{aligned}\boldsymbol{\theta}^{(t+1)} &= \boldsymbol{\theta}^{(t)} - \mathrm{Q}''(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})^{-1}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \mathrm{Q}'(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \\ &= \boldsymbol{\theta}^{(t)} - \mathrm{Q}''(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})^{-1}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \ell'(\boldsymbol{\theta}^{(t)})\end{aligned}$$

- Ascent is ensured for canonical parameters in exponential families. Backtracking can ensure ascent in other cases; inflating steps can speed convergence.

Overview