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 $L(\boldsymbol{\theta}; \boldsymbol{y}) = f(\boldsymbol{y}; \boldsymbol{\theta})$

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

$\ell(\boldsymbol{\theta}), \ell_c(\boldsymbol{\theta})$: 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

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

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.,

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

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.

Trivial 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.

Mixture of normal and uniform distributions

- Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ be an i.i.d. sample from a mixture of a normal distribution $\mathcal{N}(\mu, \sigma)$ and a uniform distribution $\mathcal{U}([-a, a])$, with pdf

$$f(y; \theta) = \pi\phi(y; \mu, \sigma) + (1 - \pi)c, \tag{1}$$

  where $\phi(\cdot; \mu, \sigma)$ is the normal pdf, $c = (2a)^{-1}$, π is the proportion of the normal distribution in the mixture and $\boldsymbol{\theta} = (\mu, \sigma, \pi)^T$ is the vector of parameters.
- Typically, the uniform distribution corresponds to outliers in the data. The proportion of outliers in the population is then $1 - \pi$.
- We want to estimate $\boldsymbol{\theta}$.

Observed and complete-data likelihoods

- Let $Z_i = 1$ if observation i is not an outlier, $Z_i = 0$ otherwise. We have $Z_i \sim \mathcal{B}(\pi)$.
- The vector $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ is the missing data.
- Observed-data likelihood:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} [\pi \phi(y_i; \mu, \sigma) + (1 - \pi)c]$$

- Complete-data likelihood:

$$\begin{aligned} L_c(\boldsymbol{\theta}) &= \prod_{i=1}^{n} f(y_i, z_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i | z_i; \mu, \sigma) f(z_i | \pi) \\ &= \prod_{i=1}^{n} \left[\phi(y_i; \mu, \sigma)^{z_i} c^{1-z_i} \pi^{z_i} (1 - \pi)^{1-z_i} \right] \end{aligned}$$

Derivation of function Q

- Complete-data log-likelihood:

$$\ell_c(\boldsymbol{\theta}) = \sum_{i=1}^{n} z_i \log \phi(y_i; \mu, \sigma) + \pi \left(n - \sum_{i=1}^{n} z_i \right) + \sum_{i=1}^{n} (z_i \log \pi + (1 - z_i) \log(1 - \pi))$$

- It is linear in the z_i. Consequently, the Q function is simply

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{n} z_i^{(t)} \log \phi(y_i; \mu, \sigma) + \pi \left(n - \sum_{i=1}^{n} z_i^{(t)} \right) + \sum_{i=1}^{n} \left(z_i^{(t)} \log \pi + (1 - z_i^{(t)}) \log(1 - \pi) \right)$$

with $z_i^{(t)} = \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[Z_i | y_i]$.

EM algorithm

E-step: compute

$$z_i^{(t)} = \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[Z_i|y_i] = \mathbb{P}_{\boldsymbol{\theta}^{(t)}}[Z_i = 1|y_i]$$
$$= \frac{\phi(y_i; \mu^{(t)}, \sigma^{(t)})\pi^{(t)}}{\phi(y_i; \mu^{(t)}, \sigma^{(t)})\pi^{(t)} + c(1-\pi^{(t)})}$$

M-step: Maximize $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$ We get

$$\pi^{(t+1)} = \sum_{i=1}^{n} z_i^{(t)}, \quad \mu^{(t+1)} = \frac{\sum_{i=1}^{n} z_i^{(t)} y_i}{\sum_{i=1}^{n} z_i^{(t)}}$$

$$\sigma^{(t+1)} = \sqrt{\frac{\sum_{i=1}^{n} z_i^{(t)} (y_i - \mu^{(t+1)})^2}{\sum_{i=1}^{n} z_i^{(t)}}}$$

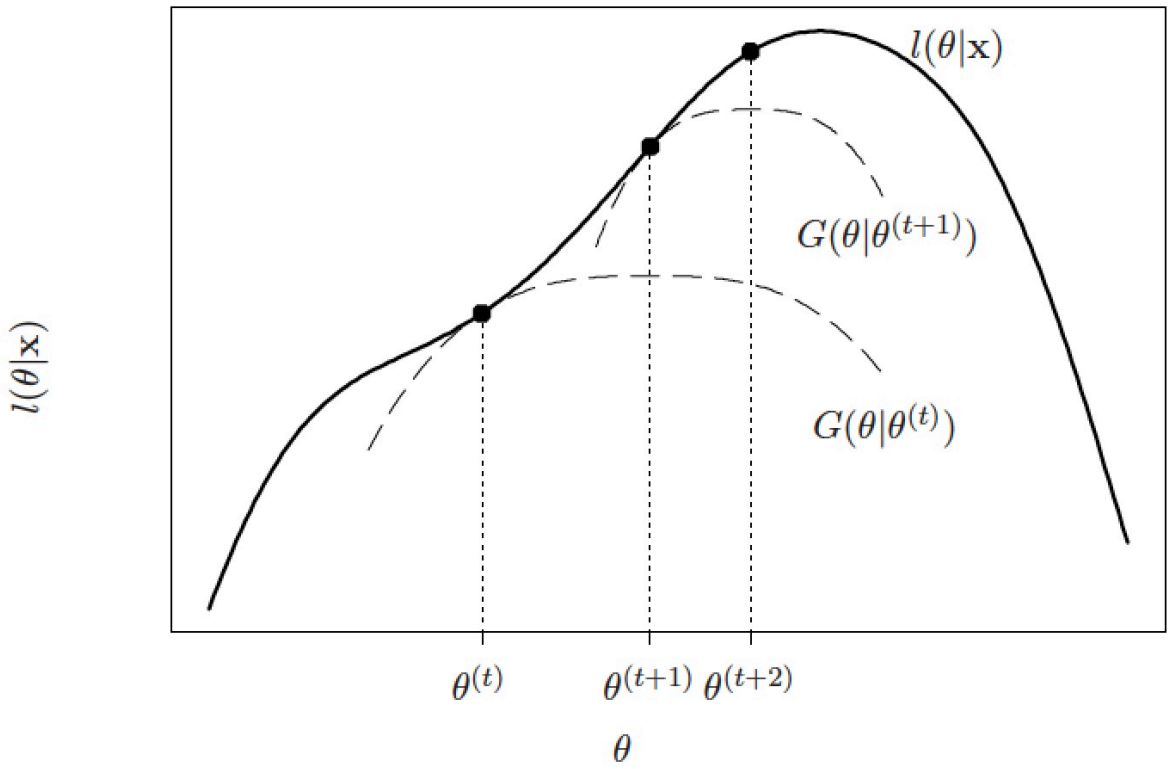
The nature of EM

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

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

- 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

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 $f(\mathbf{z}|\mathbf{y}; \boldsymbol{\theta}^{(t)})$. 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 $\hat{Q}^{(t+1)}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = \frac{1}{m^{(t)}} \sum_{j=1}^{m^{(t)}} \log f(\mathbf{X}_j^{(t)}|\boldsymbol{\theta})$.
- 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 $m^{(t)}$.

Overview

Generalized EM (GEM) algorithm

- In the original EM algorithm, $\boldsymbol{\theta}^{(t+1)}$ is a maximizer of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$, i.e.,

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

  for all $\boldsymbol{\theta}$.

- However, to ensure convergence, we only need that

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

- Any algorithm that chooses $\boldsymbol{\theta}^{(t+1)}$ at each iteration to guarantee the above condition (without maximizing $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$) is called a Generalized EM (GEM) algorithm.

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.

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.
- 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.
- ECM is a GEM algorithm, since each CM step increases Q.
- 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.

Overview

Variance of the MLE

- Let $\widehat{\boldsymbol{\theta}}$ be the MLE of $\boldsymbol{\theta}$.
- As $n \to \infty$, the limiting distribution of $\widehat{\boldsymbol{\theta}}$ is $\mathcal{N}(\boldsymbol{\theta}^*, I(\boldsymbol{\theta}^*)^{-1})$, where $\boldsymbol{\theta}^*$ is the true value of $\boldsymbol{\theta}$, and

$$I(\boldsymbol{\theta}) = \mathbb{E}[\ell'(\boldsymbol{\theta})\ell'(\boldsymbol{\theta})^T] = -\mathbb{E}[\ell''(\boldsymbol{\theta})]$$

  is the expected Fisher information matrix (the second equality holds under some regularity conditions).
- $I(\boldsymbol{\theta}^*)$ can be estimated by $I(\widehat{\boldsymbol{\theta}})$, or by $-\ell''(\widehat{\boldsymbol{\theta}}) = I_{obs}(\widehat{\boldsymbol{\theta}})$ (observed information matrix).
- Standard error estimates can be obtained by computing the square roots of the diagonal elements of $I_{obs}(\widehat{\boldsymbol{\theta}})^{-1}$.

Obtaining variance estimates

- The EM algorithms allows us to estimate $\widehat{\boldsymbol{\theta}}$, but it does not directly provide an estimate of $I(\boldsymbol{\theta}^*)$.
- Direct computation of $I(\widehat{\boldsymbol{\theta}})$ or $I_{obs}(\widehat{\boldsymbol{\theta}})$ is often difficult.
- Main methods:
  1. Louis' method
  2. SEM algorithm
  3. Bootstrap

Overview

Missing information principle

- We have seen that

$$f(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta}) = \frac{f(\boldsymbol{x};\boldsymbol{\theta})}{f(\boldsymbol{y};\boldsymbol{\theta})},$$

from which we get

$$\ell(\boldsymbol{\theta}) = \ell_c(\boldsymbol{\theta}) - \log f(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta}).$$

- Differentiating twice and negating both sides, then taking expectations over the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{y}$,

$$-\ell''(\boldsymbol{\theta}) = \mathbb{E}\left[-\ell_c''(\boldsymbol{\theta})|\boldsymbol{y}\right] - \mathbb{E}\left[-\frac{\partial^2 \log f(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}|\boldsymbol{y}\right]$$

$$\hat{\boldsymbol{\imath}}_{\mathbf{Y}}(\boldsymbol{\theta}) = \hat{\boldsymbol{\imath}}_{\mathbf{X}}(\boldsymbol{\theta}) - \hat{\boldsymbol{\imath}}_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\theta})$$

where

- $\hat{\boldsymbol{\imath}}_{\mathbf{Y}}(\boldsymbol{\theta})$ is the observed information,
- $\hat{\boldsymbol{\imath}}_{\mathbf{X}}(\boldsymbol{\theta})$ is the complete information, and
- $\hat{\boldsymbol{\imath}}_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\theta})$ is the missing information.

Louis' method

- Computing $\hat{\boldsymbol{\imath}}_{\mathbf{X}}(\boldsymbol{\theta})$ and $\hat{\boldsymbol{\imath}}_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\theta})$ is sometimes easier than computing $-\ell''(\boldsymbol{\theta})$ directly
- We can show that

$$\hat{\boldsymbol{\imath}}_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\theta}) = \mathrm{Var}[S_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\theta})],$$

where the variance is taken w.r.t. $\boldsymbol{Z}|\boldsymbol{y}$, and

$$S_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\theta}) = \frac{\partial f(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

is the conditional score.

- As the expected score is zero at $\widehat{\boldsymbol{\theta}}$, we have

$$\hat{\boldsymbol{\imath}}_{\mathbf{Z}|\mathbf{Y}}(\widehat{\boldsymbol{\theta}}) = \int S_{\mathbf{Z}|\mathbf{Y}}(\widehat{\boldsymbol{\theta}}) S_{\mathbf{Z}|\mathbf{Y}}(\widehat{\boldsymbol{\theta}})^T f(\boldsymbol{z}|\boldsymbol{y};\widehat{\boldsymbol{\theta}}) d\boldsymbol{z}$$

Monte Carlo approximation

- When they cannot be computed analytically, $\hat{\boldsymbol{\imath}}_{\mathbf{X}}(\boldsymbol{\theta})$ and $\hat{\boldsymbol{\imath}}_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\theta})$ can sometimes be approximated by Monte Carlo simulation.
- Method: generate simulated datasets $\boldsymbol{x}_j = (\boldsymbol{y}, \boldsymbol{z}_j)$, $j = 1, \ldots, N$, where $\boldsymbol{y}$ is the observed dataset, and the $\boldsymbol{z}_j$ are imputed missing datasets drawn from $f(\boldsymbol{z}|\boldsymbol{y}; \boldsymbol{\theta})$
- Then,

$$\hat{\boldsymbol{\imath}}_{\mathbf{X}}(\boldsymbol{\theta}) \approx \frac{1}{N} \sum_{j=1}^{N} -\frac{\partial^2 \log f(\boldsymbol{x}_j; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$$

and $\hat{\boldsymbol{\imath}}_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\theta})$ is approximated by the sample variance of the values

$$\frac{\partial f(\boldsymbol{z}_j|\boldsymbol{y}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

Overview

Supplemented EM (SEM) algorithm

- Let $\boldsymbol{\Psi}$ denotes the EM mapping, defined by

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\Psi}(\boldsymbol{\theta}^{(t)})$$

  having fixed point $\widehat{\boldsymbol{\theta}}$ and Jacobian matrix $\boldsymbol{\Psi}'(\boldsymbol{\theta})$ with (i,j)th element equaling $\frac{\partial \Psi_i(\boldsymbol{\theta})}{\partial \theta_j}$.

- It can be shown that

$$\boldsymbol{\Psi}'(\widehat{\boldsymbol{\theta}})^T = \hat{\boldsymbol{\imath}}_{\mathbf{Z}|\mathbf{Y}}(\widehat{\boldsymbol{\theta}})\hat{\boldsymbol{\imath}}_{\mathbf{X}}(\widehat{\boldsymbol{\theta}})^{-1}$$

- Further use of the missing information principle leads to

$$\hat{\boldsymbol{\imath}}_{\mathbf{Y}}(\widehat{\boldsymbol{\theta}})^{-1} = \hat{\boldsymbol{\imath}}_{\mathbf{X}}(\widehat{\boldsymbol{\theta}})^{-1}\left(\mathbf{I} + \boldsymbol{\Psi}'(\widehat{\boldsymbol{\theta}})^T(\mathbf{I} - \boldsymbol{\Psi}'(\widehat{\boldsymbol{\theta}})^T)^{-1}\right).$$

- SEM is numerically stable and requires little extra work.

Estimation of $\boldsymbol{\Psi}'(\widehat{\boldsymbol{\theta}})$

- Ler r_{ij} be the element (i,j) of $\boldsymbol{\Psi}'(\widehat{\boldsymbol{\theta}})$. By definition,

$$
\begin{aligned}
r_{ij} &= \frac{\partial \Psi_i(\widehat{\boldsymbol{\theta}})}{\partial \theta_j} \\
&= \lim_{\theta_j \to \widehat{\theta}_j} \frac{\Psi_i(\widehat{\theta}_1, \ldots, \widehat{\theta}_{j-1}, \theta_j, \widehat{\theta}_{j+1}, \ldots, \widehat{\theta}_p) - \Psi_i(\widehat{\boldsymbol{\theta}})}{\theta_j - \widehat{\theta}_j} \\
&= \lim_{t \to \infty} \frac{\Psi_i(\boldsymbol{\theta}^{(t)}(j)) - \Psi_i(\widehat{\boldsymbol{\theta}})}{\theta_j^{(t)} - \widehat{\theta}_j} = \lim_{t \to \infty} r_{ij}^{(t)}
\end{aligned}
$$

  where $\boldsymbol{\theta}^{(t)}(j) = (\widehat{\theta}_1, \ldots, \widehat{\theta}_{j-1}, \theta_j^{(t)}, \widehat{\theta}_{j+1}, \ldots, \widehat{\theta}_p)$, and $(\theta_j^{(t)})$, $t = 1, 2, \ldots$ is a sequence of values converging to $\widehat{\theta}_j$.

- Method: compute the $r_{ij}^{(t)}$, $t = 1, 2, \ldots$ until they stabilize to some values. Then compute $\hat{\boldsymbol{\imath}}_{\boldsymbol{Y}}(\widehat{\boldsymbol{\theta}})^{-1}$ using the previous formula.

SEM algorithm

1. Run the EM algorithm to convergence, finding $\widehat{\boldsymbol{\theta}}$.
2. Restart the algorithm from some $\boldsymbol{\theta}^{(0)}$ near $\widehat{\boldsymbol{\theta}}$. For $t = 0, 1, 2, \ldots$
   1. Take a standard E step and M step to produce $\boldsymbol{\theta}^{(t+1)}$ from $\boldsymbol{\theta}^{(t)}$.
   2. For $j = 1, \ldots, p$, define $\boldsymbol{\theta}^{(t)}(j) = (\hat{\theta}_1, \ldots, \hat{\theta}_{j-1}, \theta_j^{(t)}, \hat{\theta}_{j+1}, \ldots, \hat{\theta}_p)$ and

$$r_{ij}^{(t)} = \frac{\Psi_i(\boldsymbol{\theta}^{(t)}(j)) - \hat{\theta}_i}{\theta_j^{(t)} - \hat{\theta}_j}$$

      for $i = 1, \ldots, p$. (Recall that $\boldsymbol{\Psi}(\widehat{\boldsymbol{\theta}}) = \widehat{\boldsymbol{\theta}}$.)
   3. Stop when all $r_{ij}^{(t)}$ have converged
3. The (i, j)th element of $\boldsymbol{\Psi}'(\widehat{\boldsymbol{\theta}})$ equals $\lim_{t\to\infty} r_{ij}^{(t)}$. Use the final estimate of $\boldsymbol{\Psi}'(\widehat{\boldsymbol{\theta}})$ to get the variance.

Overview

Principle

- Consider the case of iid data $\boldsymbol{y} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)$
- If we knew the distribution of the $\boldsymbol{W}_i$, we could
  - generate many samples $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$,
  - compute the ML estimate $\widehat{\boldsymbol{\theta}}_j$ of $\boldsymbol{\theta}$ from each sample $\boldsymbol{y}_j$, and
  - estimate the variance of $\widehat{\boldsymbol{\theta}}$ by the sample variance of the estimates $\widehat{\boldsymbol{\theta}}_1, \ldots, \widehat{\boldsymbol{\theta}}_N$.
- Bootstrap principle: use the empirical distribution in place of the true distribution of the $\boldsymbol{W}_i$

Algorithm

1. Calculate $\widehat{\boldsymbol{\theta}}_{EM}$ using a suitable EM approach applied to $\boldsymbol{y} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)$. Let $j = 1$ and set $\widehat{\boldsymbol{\theta}}_j^* = \widehat{\boldsymbol{\theta}}_{EM}$.
2. Increment j. Sample pseudo-data $\boldsymbol{y}_j^* = (\boldsymbol{w}_{j1}^*, \ldots, \boldsymbol{w}_{jn}^*)$ at random from $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)$ with replacement.
3. Calculate $\widehat{\boldsymbol{\theta}}_j^*$ by applying the same EM approach to the pseudo-data $\boldsymbol{y}_j^*$
4. Stop if $j = B$ (typically, $B \geq 1000$); otherwise return to step 2.

The collection of parameter estimates $\widehat{\boldsymbol{\theta}}_1^*, \ldots, \widehat{\boldsymbol{\theta}}_B^*$ can be used to estimate the variance of $\widehat{\boldsymbol{\theta}}$,

$$\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\theta}}) = \frac{1}{B} \sum_{j=1}^{B} (\widehat{\boldsymbol{\theta}}_j^* - \overline{\widehat{\boldsymbol{\theta}}^*})(\widehat{\boldsymbol{\theta}}_j^* - \overline{\widehat{\boldsymbol{\theta}}^*})^T,$$

where $\overline{\widehat{\boldsymbol{\theta}}^*}$ is the sample mean of $\widehat{\boldsymbol{\theta}}_1^*, \ldots, \widehat{\boldsymbol{\theta}}_B^*$.

Pros and cons of the bootstrap

1. Advantages:
   - The method is very general, complex analytical derivations are avoided.
   - Allows the estimation of other aspects of the sampling distribution of $\widehat{\boldsymbol{\theta}}$, such as expectation (bias), quantiles, etc.
2. Drawback: bootstrap embeds the EM loop in a second loop of B iterations. May be computationally burdensome when the EM algorithm is slow (because, e.g., of a high proportion of missing data, or high dimensionality)