Computational statistics

Chapter 3: EM algorithm

Thierry Denœux
UniversitÃľ de technologie de CompiÃĺgne

2022-2023

EM Algorithm

- An iterative optimization strategy useful when maximizing the likelihood is difficult, but:
  - There are missing (non-observed) data
  - If the missing data were observed, maximizing the likelihood would be easy.
- Many applications in statistics and econometrics.
- Can be very simple to implement. Can reliably find an optimum through stable, uphill steps.

Overview

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

Q function

- 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

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

($f(\boldsymbol{x} \mid \boldsymbol{y}; \theta^{(t)}) = f(\boldsymbol{z} \mid \boldsymbol{y}; \theta^{(t)})$ because $\mathbf{Z}$ 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. Increment t and return to the E step unless a stopping criterion has been met; e.g.,

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

or

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

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.
- Typically, we run the algorithm several times with random initial conditions, and we keep the results of the best run.

Example: 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

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

  where $\phi(\cdot; \mu, \sigma)$ is the normal pdf, $c = (2a)^{-1}$ is a known constant, 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:

$$L_c(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i, z_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i \mid z_i; \mu, \sigma) f(z_i; \pi)$$

$$= \prod_{i=1}^{n} \left[\phi(y_i; \mu, \sigma)^{z_i} c^{1-z_i} \pi^{z_i} (1-\pi)^{1-z_i}\right]$$

Derivation of function Q

- Complete-data log-likelihood:

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

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

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

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

EM algorithm

E-step: compute

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

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

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

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

Bayesian posterior mode

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

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

- 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

1. EM algorithm
   - Description
   - Analysis
2. Some variants
   - Facilitating the E-step
   - Facilitating the M-step
3. Variance estimation
   - Louis' method
   - SEM algorithm

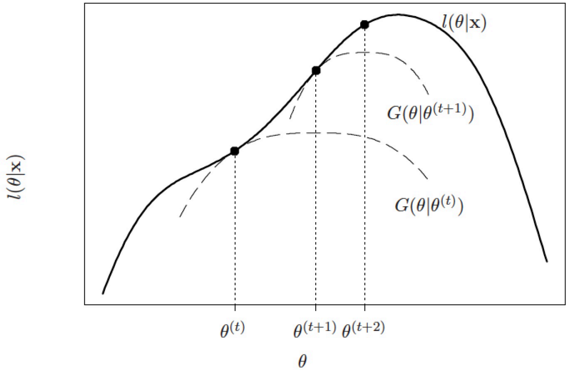
Why does it work?

- Ascent: Each M-step increases the log likelihood.
- Optimization transfer:

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

- The last two terms in $G(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$ do not depend on $\boldsymbol{\theta}$, 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 (see next slide).
- 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.

Proof

- We have

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

- Consequently,

$$\ell(\boldsymbol{\theta}) = \log f(\boldsymbol{y};\boldsymbol{\theta}) = \underbrace{\log f(\boldsymbol{x};\boldsymbol{\theta})}_{\ell_c(\boldsymbol{\theta})} - \log f(\boldsymbol{z} \mid \boldsymbol{y};\boldsymbol{\theta})$$

- Taking expectations on both sides wrt the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{Y} = \boldsymbol{y}$ and using $\boldsymbol{\theta}^{(t)}$ for $\boldsymbol{\theta}$:

$$\ell(\boldsymbol{\theta}) = Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(t)}) - \underbrace{\mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\log f(\boldsymbol{Z} \mid \boldsymbol{y};\boldsymbol{\theta}) \mid \boldsymbol{y}]}_{H(\boldsymbol{\theta},\boldsymbol{\theta}^{(t)})} \tag{2}$$

Proof: $\theta^{(t)}$ is a maximizer of $H(\theta, \theta^{(t)})$

- Now, for all $\boldsymbol{\theta} \in \Theta$,

$$H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t)}) = \mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[\log \frac{f(\boldsymbol{Z} \mid \boldsymbol{y}; \boldsymbol{\theta})}{f(\boldsymbol{Z} \mid \boldsymbol{y}; \boldsymbol{\theta}^{(t)})} \mid \boldsymbol{y}\right] \tag{3a}$$

$$\leq \log \underbrace{\mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[\frac{f(\boldsymbol{Z} \mid \boldsymbol{y}; \boldsymbol{\theta})}{f(\boldsymbol{Z} \mid \boldsymbol{y}; \boldsymbol{\theta}^{(t)})} \mid \boldsymbol{y}\right]}_{\int \frac{f(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta})}{f(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta}^{(t)})} f(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta}^{(t)}) d\boldsymbol{z}} (*) \tag{3b}$$

$$\leq \log \underbrace{\int f(\boldsymbol{z} \mid \boldsymbol{y}; \boldsymbol{\theta}) d\boldsymbol{z}}_{1} = 0 \tag{3c}$$

(*): from the concavity of the log and Jensen's inequality.

- Hence, $\boldsymbol{\theta}^{(t)}$ is a maximizer of $H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$

Proof: $\ell(\cdot)$ dominates $G(\cdot, \theta^{(t)})$

Hence, for all $\boldsymbol{\theta} \in \Theta$,

$$H(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t)}) \geq H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$$

$$Q(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t)}) - \ell(\boldsymbol{\theta}^{(t)}) \geq Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) - \ell(\boldsymbol{\theta})$$

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

Proof: G is tangent to ℓ at $\boldsymbol{\theta}^{(t)}$

- As $\boldsymbol{\theta}^{(t)}$ maximizes $H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) - \ell(\boldsymbol{\theta})$, we have

$$H'(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} = Q'(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} - \ell'(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} = 0,$$

  so

$$Q'(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} = \ell'(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}.$$

- Consequently, as $G(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) + \text{cst}$,

$$G'(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} = Q'(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} = \ell'(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}.$$

Proof: monotonicity

- From (2),

$$\ell(\boldsymbol{\theta}^{(t+1)}) - \ell(\boldsymbol{\theta}^{(t)}) = \underbrace{Q(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t)})}_{A} - \left[\underbrace{H(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t)})}_{B}\right]$$

- $A \geq 0$ because $\boldsymbol{\theta}^{(t+1)}$ is a maximizer of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$, and $B \leq 0$ because, from (3), $\boldsymbol{\theta}^{(t)}$ is a maximizer of $H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$.
- Hence,

$$\boxed{\ell(\boldsymbol{\theta}^{(t+1)}) \geq \ell(\boldsymbol{\theta}^{(t)})}$$

Overview

Overview

Monte Carlo EM (MCEM)

- Sometimes, the conditional expectation of $\ell_c(\boldsymbol{\theta})$ given $\boldsymbol{y}$ cannot be easily computed analytically in the E step.
- Approach: randomly generate sets of missing values according to the conditional distribution $f(\mathbf{z}|\mathbf{y};\boldsymbol{\theta}^{(t)})$, and replace the expectation by an average over generated data sets.

Monte Carlo EM (MCEM)

- Replace the t-th 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

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

- Then $\widehat{Q}^{(t+1)}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$ is a Monte Carlo estimate of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$.
- The M step is modified to maximize $\widehat{Q}^{(t+1)}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$.

Remarks

- It is advised to increase $m^{(t)}$ as iterations progress to reduce the Monte Carlo variability of $\widehat{Q}$.
- 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 up convergence.

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

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

  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 algorithm 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. Supplemented EM (SEM) algorithm
  3. Bootstrap (to be studied in Chapter 6)

Overview

Missing information principle

- We have seen that

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

from which we get

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

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

$$\underbrace{-\ell''(\boldsymbol{\theta})}_{\hat{\imath}_{\mathbf{Y}}(\boldsymbol{\theta})} = \underbrace{\mathbb{E}_{\boldsymbol{\theta}}\left[-\ell_c''(\boldsymbol{\theta}) \mid \boldsymbol{y}\right]}_{\hat{\imath}_{\mathbf{X}}(\boldsymbol{\theta})} - \underbrace{\mathbb{E}_{\boldsymbol{\theta}}\left[-\frac{\partial^2 \log f(\boldsymbol{z} \mid \boldsymbol{y}; \boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} \mid \boldsymbol{y}\right]}_{\hat{\imath}_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\theta})}$$

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

$$\hat{\boldsymbol{\imath}}_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\theta}) = \text{Var}\left[S_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\theta}) \mid \boldsymbol{y}\right],$$

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

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

is the conditional score.

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

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

Monte Carlo approximation

- When $\hat{\boldsymbol{\imath}}_{\mathbf{X}}(\boldsymbol{\theta})$ and $\hat{\boldsymbol{\imath}}_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\theta})$ cannot be computed analytically, they 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

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

Overview

EM mapping

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

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

- From the convergence of EM, $\widehat{\boldsymbol{\theta}}$ is a fixed point:

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{\Psi}(\widehat{\boldsymbol{\theta}}).$$

- The Jacobian matrix of $\boldsymbol{\Psi}$ is the $p \times p$ matrix

$$\boldsymbol{\Psi}'(\boldsymbol{\theta}) = \left(\frac{\partial \Psi_i(\boldsymbol{\theta})}{\partial \theta_j}\right).$$

- It can be shown that

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

Using $\boldsymbol{\Psi}'(\theta)$ for variance estimation

- From the missing information principle,

$$\begin{aligned}\hat{\imath}_{\mathbf{Y}}(\widehat{\theta}) &= \hat{\imath}_{\mathbf{X}}(\widehat{\theta}) - \hat{\imath}_{\mathbf{Z}|\mathbf{Y}}(\widehat{\theta}) \\ &= \left[\mathbf{I} - \hat{\imath}_{\mathbf{Z}|\mathbf{Y}}(\widehat{\theta})\hat{\imath}_{\mathbf{X}}(\widehat{\theta})^{-1}\right]\hat{\imath}_{\mathbf{X}}(\widehat{\theta}) \\ &= \left[\mathbf{I} - \boldsymbol{\Psi}'(\widehat{\theta})^{T}\right]\hat{\imath}_{\mathbf{X}}(\widehat{\theta}).\end{aligned}$$

- Hence,

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

Using $\boldsymbol{\Psi}'(\boldsymbol{\theta})$ for variance estimation (continued)

- From the equality

$$(\mathbf{I}-\mathbf{P})^{-1} = (\mathbf{I}-\mathbf{P}+\mathbf{P})(\mathbf{I}-\mathbf{P})^{-1} = \mathbf{I}+\mathbf{P}(\mathbf{I}-\mathbf{P})^{-1},$$

we get

$$\hat{\boldsymbol{\imath}}_{\mathbf{Y}}(\widehat{\boldsymbol{\theta}})^{-1} = \hat{\boldsymbol{\imath}}_{\mathbf{X}}(\widehat{\boldsymbol{\theta}})^{-1}\left\{\mathbf{I}+\boldsymbol{\Psi}'(\widehat{\boldsymbol{\theta}})^{T}\left[\mathbf{I}-\boldsymbol{\Psi}'(\widehat{\boldsymbol{\theta}})^{T}\right]^{-1}\right\} \quad (4)$$

- This result is appealing in that it expresses the desired covariance matrix as the complete-data covariance matrix plus an incremental matrix that takes account of the uncertainty attributable to the missing data.

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

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

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

  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 (4).

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 treating it as the current estimate of $\boldsymbol{\theta}$, run one iteration of EM to obtain $\Psi(\boldsymbol{\theta}^{(t)}(j))$.
        - Obtain the ratio
        $$r_{ij}^{(t)} = \frac{\Psi_i(\boldsymbol{\theta}^{(t)}(j)) - \hat{\theta}_i}{\theta_j^{(t)} - \hat{\theta}_j}$$
        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.