EM ALGORITHM

HAN XIAO

0. Review of Multivariate Normal Distribution

Suppose $X \in \mathbb{R}^p$ is a random vector with mean $\mu \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$. Then the following are three equivalent definitions of *multivariate normal distribution*.

(i) For any $\alpha \in \mathbb{R}^p$, $\alpha' X$ has the univariate normal distribution.
(ii) There exist a matrix $\boldsymbol{A} \in \mathbb{R}^{p \times r}$, where $r = \mathrm{rank}(\Sigma)$, and a random vector $Z = (Z_1, \ldots, Z_r)$, where $Z_1, \ldots, Z_r$ are independent standard univariate normal random variables, such that $X = \mu + \boldsymbol{A}Z$.
(iii) The moment generating function (MGF) of X is $M_X(t) = \mathbb{E}(e^{t'X}) = \exp(\mu' t + t'\Sigma t/2)$ for all $t \in \mathbb{R}^p$.

When X satisfies the definitions above, we write $X \sim \mathcal{N}_p(\mu, \Sigma)$. The subscript p indicates that X is p-dimensional, and we sometimes suppress it. Using any one of these definitions, we have the following result.

- If $X \sim \mathcal{N}_p(0, \Sigma)$, $\boldsymbol{B} \in \mathbb{R}^{m \times p}$, $\nu \in \mathbb{R}^m$, and $Y = \boldsymbol{B}X + \nu$, then $Y \sim \mathcal{N}_m(\boldsymbol{B}\mu + \nu, \boldsymbol{B}\Sigma\boldsymbol{B}')$.

If $X \sim \mathcal{N}_p(\mu, \Sigma)$ and Σ is nonsingular, then the density function of X is given by

$$f_X(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\}, \quad \text{for } x \in \mathbb{R}^p.$$

Suppose $X \sim \mathcal{N}(\mu, \Sigma)$. We partition X into two subvectors X_1 and X_2, and partition μ and Σ accordingly

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right).$$

- As a special case of the previous result, we know $X_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$ and $X_2 \sim \mathcal{N}(\mu_2, \Sigma_{22})$.
- X_1 and X_2 are independent if and only if $\Sigma_{12} = 0$. (WARNING. This is true only when X is normal. The normality of only X_1 and X_2 is not sufficient.)
- The conditional distribution of X_2 given X_1 is still normal

  $$(X_2 | X_1 = x_1) \sim \mathcal{N}\left(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)$$

  While the conditional mean depends on the value of X_1, the conditional covariance matrix does not!

1. Gaussian Mixtures

Consider a constructive mixture model where there is a latent variable Z with the multinomial distribution: $P(Z = k) = \pi_k$ for $1 \le k \le K$; and given $Z = k$, X has the density $p_k(X; \theta_k)$ governed by a set of parameters denoted by θ_k. The marginal density of X is $\sum_{k=1}^K \pi_k p_k(X; \theta_k)$, which is governed by the collection of parameters $\theta := \{\pi_1, \theta_1, \ldots, \pi_K, \theta_K\}$. We say X has a *mixture distribution*. In particular, if each p_k is a normal density, then X is called a *Gaussian mixture*.

Alternatively, let $Z = (Z_1, \ldots, Z_K)'$ be a K-dimensional random vector whose distribution is given by $P(Z = e_k) = \pi_k$ for $1 \le k \le K$, where $e_k \in \mathbb{R}^K$ is the vector whose k-th entry is one and all other entries are zero. Let X be a random variable such that given $Z = k$, X has the density $p_k(X; \theta_k)$. Then the joint density of (X, Z) is given by

$$f(X, Z) = \prod_{k=1}^K \pi_k^{Z_k} [p_k(X; \theta_k)]^{Z_k};$$

and the marginal density of X is also $\sum_{k=1}^K \pi_k p_k(X; \theta_k)$.

This is a supplementary reading for FSRM588. Last updated on October 9, 2012.

In practice, Z is not observable, and we only have a set of observations $x_1, \ldots, x_N$ on X. Consider a situation where we want to find the MLE of θ, but likelihood $p(\boldsymbol{X};\theta)$ is difficult to maximize. In many cases, if $\boldsymbol{Z}$ were observed, the log likelihood function based on the "complete" data $(x_1, z_1), \ldots, (x_N, z_N)$,

$$\ell(\boldsymbol{X}, \boldsymbol{Z}; \theta) = \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \left[\log(\pi_k) + \log p_k(x_i; \theta_k) \right], \tag{1.1}$$

is much easier to optimize. This is, however, impractical because $\boldsymbol{Z}$ is not observed. We introduce an iterative algorithm to resolve this issue. The idea is heuristic. Suppose there is an initial estimate θ^{old}. Although $\boldsymbol{Z}$ is not observed, we can compute the conditional expectation of $\boldsymbol{Z}$ given $\boldsymbol{X}$, holding the parameter at θ^{old}, *i.e.*

$$\gamma(z_{ik}) = \mathbb{E}(z_{ik} | \boldsymbol{X}; \theta^{\text{old}}); \tag{1.2}$$

and then consider the surrogate function

$$Q(\boldsymbol{X}; \theta) = \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma(z_{ik}) \left[\log(\pi_k) + \log p_k(x_i; \theta_k) \right]. \tag{1.3}$$

It shares a resemblance with $\ell(\boldsymbol{X}, \boldsymbol{Z}; \theta)$, and might be easy to maximize as well. If this is the case, we can obtain an updated estimate

$$\theta^{\text{new}} = \arg\max_{\theta} Q(\boldsymbol{X}; \theta | \theta^{\text{old}}). \tag{1.4}$$

By iterating (1.2) and (1.4), hopefully the estimate will converge to the MLE $\hat{\theta}$.

Let us use Gaussian mixtures for illustration. Assume for each k, p_k is the density of $\mathcal{N}(\mu_k, \Sigma_k)$. Here $\theta_k = \{\mu_k, \Sigma_k\}$. The likelihood function for the complete data is

$$\ell(\boldsymbol{X}, \boldsymbol{Z}; \theta) = \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \left[\log(\pi_k) - \frac{p}{2} \log(2\pi) - \frac{1}{2} \log|\Sigma_k| - \frac{1}{2} (x_i - \mu_k)' \Sigma_k^{-1/2} (x_i - \mu_k) \right].$$

For the expectation step (1.2), we compute

$$\gamma(z_{ik}) = \frac{\pi_k^{\text{old}} p_k(x_i; \mu_k^{\text{old}}, \Sigma_k^{\text{old}})}{\sum_{l=1}^{M} \pi_l^{\text{old}} p_l(x_i; \mu_l^{\text{old}}, \Sigma_l^{\text{old}})}.$$

For the maximization step (1.4), we have

$$\begin{aligned}
\pi_k^{\text{new}} &= \frac{\sum_{i=1}^{N} \gamma(z_{ik})}{N}; \\
\mu_k^{\text{new}} &= \frac{\sum_{i=1}^{N} \gamma(z_{ik}) x_i}{\sum_{i=1}^{N} \gamma(z_{ik})}; \\
\Sigma_k^{\text{new}} &= \frac{\sum_{i=1}^{N} \gamma(z_{ik}) \left(x_i - \mu_k^{\text{new}}\right) \left(x_i - \mu_k^{\text{new}}\right)'}{\sum_{i=1}^{N} \gamma(z_{ik})}.
\end{aligned}$$

The algorithm stops when the difference between two successive estimates is smaller than some pre-specified threshold, or the increase in the value of the log likelihood function between two iterations is less than some tolerance parameter.

2. EM Algorithm in General

We shall give some hints on why the algorithm introduced heuristically in the preceding section does maximize the log likelihood function. Consider a general situation in which the observed data $\boldsymbol{X}$ is augmented by some hidden variables $\boldsymbol{Z}$ to form the "complete" data, where $\boldsymbol{Z}$ can be either real missing data or artificially but cleverly constructed variables. Assume the joint density $p(\boldsymbol{X}, \boldsymbol{Z}; \theta)$ is governed by a set of parameters abbreviated by θ. The MLE of θ based on the observed data $\boldsymbol{X}$ is the solution of the following optimization problem

$$\hat{\theta} = \arg\max_{\theta} \ell(\boldsymbol{X}; \theta), \quad \text{where } \ell(\boldsymbol{X}; \theta) = \log p(\boldsymbol{X}; \theta) \text{ and } p(\boldsymbol{X}; \theta) = \int p(\boldsymbol{X}, \boldsymbol{Z}; \theta)\, d\boldsymbol{Z}. \tag{2.1}$$

Let $q(\boldsymbol{Z};\theta^{\text{old}})$ be a density of $\boldsymbol{Z}$ which is governed by an initial estimate θ^{old}. By Jensen's inequality

$$\ell(\boldsymbol{X};\theta) = \log\left[\int \frac{p(\boldsymbol{X},\boldsymbol{Z};\theta)}{q(\boldsymbol{Z};\theta^{\text{old}})} q(\boldsymbol{Z};\theta^{\text{old}})\, d\boldsymbol{Z}\right] \geq \int \log\left[\frac{p(\boldsymbol{X},\boldsymbol{Z};\theta)}{q(\boldsymbol{Z};\theta^{\text{old}})}\right] q(\boldsymbol{Z};\theta^{\text{old}})\, d\boldsymbol{Z} =: Q(\boldsymbol{X};\theta|\theta^{\text{old}}). \tag{2.2}$$

The function $Q(\boldsymbol{X};\theta|\theta^{\text{old}})$ is called a *minorization* of $\ell(\boldsymbol{X};\theta)$. If the condition

$$\arg\min_{\theta}\left[\ell(\boldsymbol{X};\theta) - Q(\boldsymbol{X};\theta|\theta^{\text{old}})\right] = \theta^{\text{old}} \tag{2.3}$$

holds, and

$$\theta^{\text{new}} = \arg\max_{\theta} Q(\boldsymbol{X};\theta|\theta^{\text{old}}); \tag{2.4}$$

then it holds that

$$\begin{aligned} \ell(\boldsymbol{X};\theta^{\text{new}}) &= \left[\ell(\boldsymbol{X};\theta^{\text{new}}) - Q(\boldsymbol{X};\theta^{\text{new}}|\theta^{\text{old}})\right] + Q(\boldsymbol{X};\theta^{\text{new}}|\theta^{\text{old}}) \\ &\geq \left[\ell(\boldsymbol{X};\theta^{\text{old}}) - Q(\boldsymbol{X};\theta^{\text{old}}|\theta^{\text{old}})\right] + Q(\boldsymbol{X};\theta^{\text{old}}|\theta^{\text{old}}) = \ell(\boldsymbol{X};\theta^{\text{old}}). \end{aligned} \tag{2.5}$$

The message here is that when $\ell(\boldsymbol{X};\theta)$ is difficult to maximize, but the surrogate function $Q(\boldsymbol{X};\theta|\theta^{\text{old}})$ is significantly easier to maximize; starting with an initial estimate θ^{old}, through a minorization step (2.2) and a maximization step (2.4), we can obtain a new estimate θ^{new} which necessarily increases the value of the log likelihood function $\ell(\boldsymbol{X};\theta)$. The procedure we just described is under the general MM (minorize-maximize) framework (Lange et al., 2000).

We still need a specification of $q(\boldsymbol{Z};\theta^{\text{old}})$ such that (2.3) holds. In the EM algorithm, one choose the conditional density of $\boldsymbol{Z}$ given $\boldsymbol{X}$, evaluated at θ^{old}, *i.e.*

$$q(\boldsymbol{Z};\theta^{\text{old}}) = \frac{p(\boldsymbol{X},\boldsymbol{Z};\theta^{\text{old}})}{p(\boldsymbol{X};\theta^{\text{old}})}. \tag{2.6}$$

The condition (2.3) holds for this choice because

$$\ell(\boldsymbol{X};\theta^{\text{old}}) - Q(\boldsymbol{X};\theta^{\text{old}}|\theta^{\text{old}}) = 0.$$

Now let us take a new look at the minorization step (2.2) and maximization step (2.4). Since

$$Q(\boldsymbol{X};\theta|\theta^{\text{old}}) = \int \log[p(\boldsymbol{X},\boldsymbol{Z};\theta)]\cdot q(\boldsymbol{Z};\theta^{\text{old}})\, d\boldsymbol{Z} - \int \log[q(\boldsymbol{Z};\theta^{\text{old}})]\cdot q(\boldsymbol{Z};\theta^{\text{old}})\, d\boldsymbol{Z},$$

it suffices to consider the function

$$\tilde{Q}(\boldsymbol{X};\theta|\theta^{\text{old}}) = \int \log[p(\boldsymbol{X},\boldsymbol{Z};\theta)]\cdot q(\boldsymbol{Z};\theta^{\text{old}})\, d\boldsymbol{Z}; \tag{2.7}$$

and then find the maximizer

$$\theta^{\text{new}} = \arg\max_{\theta} \tilde{Q}(\boldsymbol{X};\theta|\theta^{\text{old}}). \tag{2.8}$$

In (2.7), $\tilde{Q}(\boldsymbol{X};\theta|\theta^{\text{old}})$ is the expectation of the log likelihood (for complete data) taken over the conditional distribution of $\boldsymbol{Z}$ given $\boldsymbol{X}$, so it is called a *E-step*, where "E" stands for expectation. The maximization step (2.8) is called a *M-step*. The algorithm is thus called *EM (expectation-maximization) algorithm*. The right panel of Figure 1 illustrates the behavior of the EM algorithm* .

3. EM for Factor Models

If F and E are jointly normal, then (F, X) is jointly normal with

$$\begin{pmatrix} F \\ X \end{pmatrix} \sim \mathcal{N}\left\{\begin{pmatrix} 0 \\ \mu \end{pmatrix}, \begin{pmatrix} \boldsymbol{I} & \boldsymbol{\beta}' \\ \boldsymbol{\beta} & \Psi + \boldsymbol{\beta}\boldsymbol{\beta}' \end{pmatrix}\right\}. \tag{3.1}$$

If both F and X were observable, the likelihood function based on $(x_1, f_1), \ldots, (x_N, f_N)$ would be

$$\prod_{i=1}^{N}\left\{\frac{1}{(2\pi)^{p/2}|\Psi|^{1/2}} \exp\left[-\frac{1}{2}(x_i - \mu - \boldsymbol{\beta} f_i)'\Psi^{-1}(x_i - \mu - \boldsymbol{\beta} f_i)\right] \times \frac{1}{(2\pi)^{m/2}} \exp\left(-\frac{1}{2}f_i' f_i\right)\right\}.$$

*It can be shown that the surrogate $Q(\boldsymbol{X};\theta|\theta^{\text{old}})$ and the log likelihood $\ell(\boldsymbol{X};\theta)$ have the same gradient at the point θ^{old}.

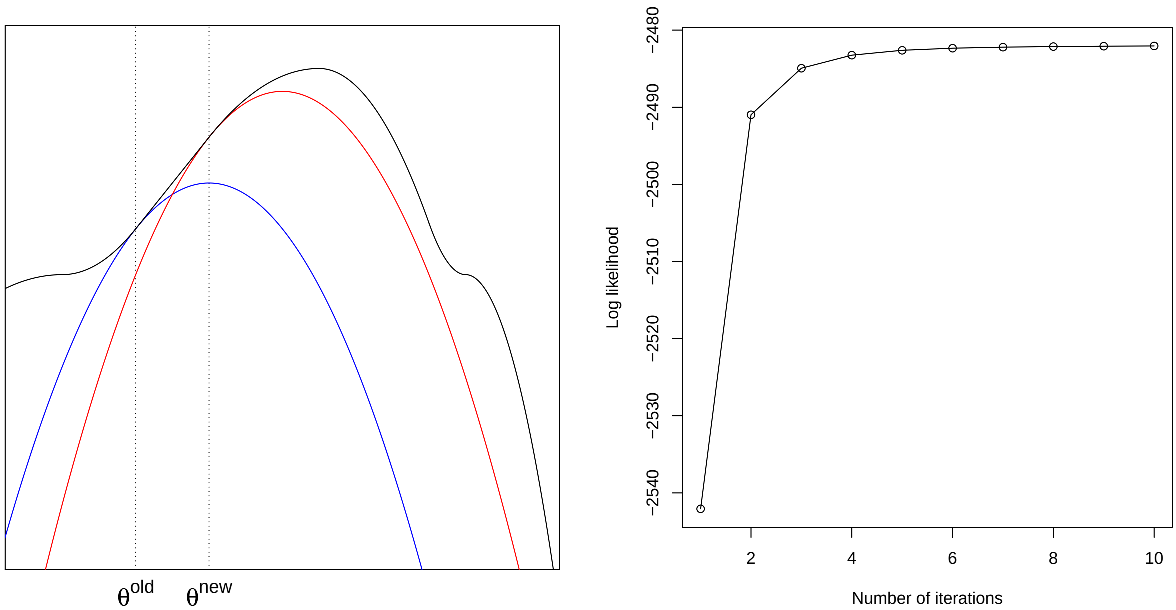

FIGURE 1. The left panel provides an illustration of the EM algorithm. The black curve gives the log likelihood $\ell(\boldsymbol{X};\theta)$. The blue curve is the surrogate function $Q(\boldsymbol{X};\theta|\theta^{\mathrm{old}})$ which makes an tangential contact with $\ell(\boldsymbol{X};\theta)$ at the point θ^{old}. The maximizer of $Q(\boldsymbol{X};\theta|\theta^{\mathrm{old}})$ is θ^{new}. The red curve is the surrogate function $Q(\boldsymbol{X};\theta|\theta^{\mathrm{new}})$ at θ^{new}. The right panel depicts the increase of the log likelihood as EM algorithm proceeds.

It suffices to consider the following part of the log likelihood

$$J(\boldsymbol{X},\boldsymbol{F};\theta) = -\frac{N}{2}\log|\Psi| - \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu-\boldsymbol{\beta}f_i)'\Psi^{-1}(x_i-\mu-\boldsymbol{\beta}f_i),$$

where θ denotes the set of parameters $\theta=\{\mu,\boldsymbol{\beta},\Psi\}$. In the E-step, we evaluate the expectation of $J(\boldsymbol{X},\boldsymbol{F};\theta)$ over a distribution $q(\boldsymbol{F};\theta^{\mathrm{old}})$ of $\boldsymbol{F}$. Denote this expectation by $\mathbb{E}_q$. Assume

$$\mathbb{E}_q\bar{f} := \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_q f_i = 0. \tag{3.2}$$

We will explain why we can make such an assumption later. Under it we have†

$$\begin{aligned}
\mathbb{E}_q J(\boldsymbol{X},\boldsymbol{F};\theta) &= \mathbb{E}_q\sum_{j=1}^{p}\left\{-\frac{1}{2}\log(\sigma_j^2) - \frac{1}{2N\sigma_j^2}\sum_{i=1}^{N}(x_{ij}-\bar{x}_j-f_i'\beta_j)^2 - \frac{1}{2\sigma_j^2}(\mu_j-\bar{x}_j)^2\right\} && (3.3)\\
&= -\frac{1}{2}\log|\Psi| - \frac{1}{2N}\mathbb{E}_q\left\|(\boldsymbol{X}^c-\boldsymbol{F}\boldsymbol{\beta}')\Psi^{-1/2}\right\|_F^2 - \frac{1}{2}(\mu-\bar{x})\Psi^{-1}(\mu-\bar{x})'\\
&= -\frac{1}{2}\log|\Psi| - \frac{1}{2N}\operatorname{tr}\left[\mathbb{E}_q(\boldsymbol{X}^c-\boldsymbol{F}\boldsymbol{\beta}')'(\boldsymbol{X}^c-\boldsymbol{F}\boldsymbol{\beta}')\Psi^{-1}\right] - \frac{1}{2}(\mu-\bar{x})\Psi^{-1}(\mu-\bar{x})'\\
&= -\frac{1}{2}\log|\Psi| - \frac{1}{2N}\operatorname{tr}\left[\mathbb{E}_q(\boldsymbol{\beta}\boldsymbol{C}_{ff}\boldsymbol{\beta}'-\boldsymbol{C}_{xf}\boldsymbol{\beta}'-\boldsymbol{\beta}\boldsymbol{C}_{xf}'+\boldsymbol{C}_{xx})\Psi^{-1}\right] - \frac{1}{2}(\mu-\bar{x})\Psi^{-1}(\mu-\bar{x})', && (3.4)
\end{aligned}$$

†We can rewrite

$$\sum_{i=1}^{N}(x_{ij}-\mu_j-f_i'\beta_j)^2 = \sum_{i=1}^{N}(x_{ij}-\bar{x}_j-f_i'\beta_j)^2 + N(\mu_j-\bar{x})^2 - 2(\mu_j-\bar{x})\sum_{i=1}^{N}(x_{ij}-\bar{x}_j-\beta_j'f_i).$$

The conditional expectation of the third term on the right hand side would be 0 provided that $\sum_{i=1}^{N}\mathbb{E}_q f_i = 0$.

where

$$\boldsymbol{C}_{xx} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})' = \frac{1}{N}(\boldsymbol{X}^c)'\boldsymbol{X}^c,$$
$$\boldsymbol{C}_{xf} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})f_i' = \frac{1}{N}(\boldsymbol{X}^c)'\boldsymbol{F},$$
$$\boldsymbol{C}_{ff} = \frac{1}{N}\sum_{i=1}^{N}f_i f_i' = \frac{1}{N}\boldsymbol{F}'\boldsymbol{F}.$$

The MLE of μ must be $\bar{x}$. With an initial estimate $\theta^{\text{old}} = \{\bar{x}, \boldsymbol{\beta}^{\text{old}}, \Psi^{\text{old}}\}$, we take $q(\boldsymbol{Z};\theta^{\text{old}})$ as the conditional density of $\boldsymbol{Z}$ given $\boldsymbol{X}$, holding the parameter at θ^{old}. Using (3.1), we have

$$\begin{aligned}
\boldsymbol{C}^*_{xf} &:= \mathbb{E}_q\boldsymbol{C}_{xf} = \mathbb{E}(\boldsymbol{C}_{xf}|\boldsymbol{X};\theta^{\text{old}}) = \boldsymbol{C}_{xx}\left[\Psi^{\text{old}} + \boldsymbol{\beta}^{\text{old}}(\boldsymbol{\beta}^{\text{old}})'\right]^{-1}\boldsymbol{\beta}^{\text{old}};\\
\boldsymbol{C}^*_{ff} &:= \mathbb{E}_q\boldsymbol{C}_{ff} = \mathbb{E}(\boldsymbol{C}_{ff}|\boldsymbol{X};\theta^{\text{old}}) = \boldsymbol{I} - (\boldsymbol{\beta}^{\text{old}})'(\Sigma^{\text{old}})^{-1}\boldsymbol{\beta}^{\text{old}} + (\boldsymbol{\beta}^{\text{old}})'(\Sigma^{\text{old}})^{-1}\boldsymbol{C}_{xx}(\Sigma^{\text{old}})^{-1}\boldsymbol{\beta}^{\text{old}}\\
&\qquad\qquad\qquad\qquad\qquad\qquad\quad = \boldsymbol{I} + (\boldsymbol{\beta}^{\text{old}})'(\Sigma^{\text{old}})^{-1}(\boldsymbol{C}_{xx} - \Sigma^{\text{old}})(\Sigma^{\text{old}})^{-1}\boldsymbol{\beta}^{\text{old}}.
\end{aligned}$$

From (3.3), for the M-step we should compute

$$\mu_j^{\text{new}} = \bar{x}_j, \quad \beta_j^{\text{new}} = N^{-1}(C^*_{ff})^{-1}\mathbb{E}_q(\boldsymbol{F}'\boldsymbol{x}_j^c), \quad (\sigma_j^2)^{\text{new}} = N^{-1}(\boldsymbol{x}_j^c)'\boldsymbol{x}_j^c - N^{-2}\mathbb{E}_q[(\boldsymbol{x}_j^c)'\boldsymbol{F}](\boldsymbol{C}^*_{ff})^{-1}\mathbb{E}_q(\boldsymbol{F}'\boldsymbol{x}_j^c);$$

which can be written in the following matrix form

$$\mu^{\text{new}} = \bar{x}, \quad \boldsymbol{\beta}^{\text{new}} = \boldsymbol{C}^*_{xf}(\boldsymbol{C}^*_{ff})^{-1}, \quad \Psi^{\text{new}} = \operatorname{diag}\left[\boldsymbol{C}^*_{xx} - \boldsymbol{C}^*_{xf}(\boldsymbol{C}^*_{ff})^{-1}(\boldsymbol{C}^*_{xf})'\right], \tag{3.5}$$

where diag$(\cdot)$ extracts the diagonal elements of a matrix to form a diagonal matrix. It can be shown (Anderson, 2003) that at the MLE $\hat{\theta} = \{\bar{x}, \hat{\boldsymbol{\beta}}, \hat{\Psi}\}$

$$\hat{\Sigma}^{-1}(\boldsymbol{C}_{xx} - \hat{\Sigma})\hat{\Sigma}^{-1}\boldsymbol{\beta} = \boldsymbol{0}, \quad \text{and} \quad \operatorname{diag}(\hat{\Psi} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}') = \operatorname{diag}(\boldsymbol{C}_{xx}),$$

where $\hat{\Sigma} = \hat{\Psi} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'$; and it follows that $\hat{\theta}$ is a stable point[‡] of the EM algorithm.

Now we shall go back to the assumption (3.2). In the EM algorithm, the conditional expectation of f_i given $\boldsymbol{X}$ is $\mathbb{E}(f_i|\boldsymbol{X};\theta^{\text{old}}) = (\boldsymbol{\beta}^{\text{old}})'(\Sigma^{\text{old}})^{-1}(x_i - \bar{x})$, and hence $N^{-1}\sum_{i=1}^{N}\mathbb{E}(f_i|\boldsymbol{X};\theta^{\text{old}}) = 0$, leading to (3.2).

After the algorithm converges to the final estimate $\hat{\theta} = \{\bar{x}, \hat{\boldsymbol{\beta}}, \hat{\Psi}\}$, one can further rotate $\hat{\boldsymbol{\beta}}$ so that $\hat{\boldsymbol{\beta}}'\hat{\Psi}^{-1}\hat{\boldsymbol{\beta}}$ is diagonal. During the iterations of the algorithm, such rotations are not necessary[§]. Rubin and Thayer (1982) discussed EM algorithms under other identifiability constraints.

Example 3.1. This is a continuation of Example 3.1 of Notes 04. We run the EM algorithm for a 3-factor model, using the estimates given by PCA as initial values. The right panel of Figure 1 depicts how the log likelihood increases as the algorithm proceeds. The increase becomes negligible after 10 iterations. The estimated correlation matrix is identical to the one given by the `factanal()` function of `R`.

References

T. W. Anderson. *An introduction to multivariate statistical analysis.* Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, third edition, 2003.

Christopher M. Bishop. *Pattern recognition and machine learning.* Springer, New York, 2006.

Kenneth Lange, David R. Hunter, and Ilsoon Yang. Optimization transfer using surrogate objective functions (with discussion). *J. Comput. Graph. Statist.*, 9(1):1–59, 2000.

D. B. Rubin and D. T. Thayer. EM algorithms for ML factor analysis. *Psychometrika*, 47(1):69–76, 1982.

[‡]It holds that $\mathbb{E}(\boldsymbol{C}_{xf}|\boldsymbol{X},\hat{\theta}) = \hat{\boldsymbol{\beta}}$ and $\mathbb{E}(\boldsymbol{C}_{ff}|\boldsymbol{X},\hat{\theta}) = \boldsymbol{I}$. So if we use $\hat{\theta}$ as the initial point for the EM algorithm, then after one iteration $\boldsymbol{\beta}^{\text{new}} = \hat{\boldsymbol{\beta}}$ and $\Psi^{\text{new}} = \operatorname{diag}(\boldsymbol{C}_{xx} - \boldsymbol{\beta}\boldsymbol{\beta}') = \hat{\Psi}$.

[§]During an iteration, one can perform $\boldsymbol{\beta}^{\text{old}}Q$ for some orthogonal matrix Q so that $Q'(\boldsymbol{\beta}^{\text{old}})'(\Psi^{\text{old}})^{-1}\boldsymbol{\beta}^{\text{old}}Q$ is orthogonal, and then the new estimate will be $\boldsymbol{\beta}^{\text{new}}Q$, which is equivalent to $\boldsymbol{\beta}^{\text{new}}$ is the sense $\boldsymbol{\beta}^{\text{new}}Q(\boldsymbol{\beta}^{\text{new}}Q)' = \boldsymbol{\beta}^{\text{new}}(\boldsymbol{\beta}^{\text{new}})'$.