Week 6: The EM-Algorithm

MATH-517 Statistical Computation and Visualization

Linda Mhalla

2024-10-18

Section 1

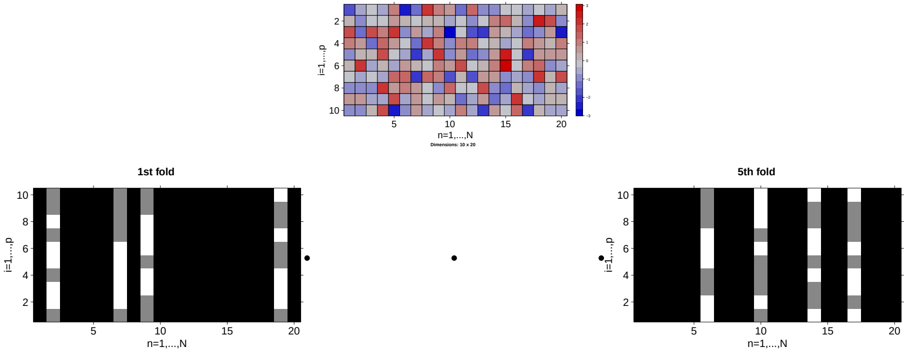
Motivation From Last Week

CV for PCA Repaired

Assume that data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ are i.i.d. realizations of $X \sim \mathcal{N}(\mu, \Sigma)$

- split data into K folds: $J_1, \ldots, J_K$
- **for** $k = 1, \ldots, K$
  - estimate μ and Σ empirically using all but the k-th fold J_k, and truncate Σ to be rank-r
  - **for** $n \in J_k$
    - split $\mathbf{x}_n$ into a "missing" part $\mathbf{x}^{miss}$ that will be used for validation and an "observed" part $\mathbf{x}^{obs}$
    - predict $\mathbf{x}_n^{miss}$ from $\mathbf{x}_n^{obs}$ as discussed on the previous slide
  - **end for**
  - calculate $Err_k(r) = \sum_{n \in J_k} \|(\mathbf{x}_n^{obs}, \mathbf{x}_n^{miss})^\top - (\mathbf{x}_n^{obs}, \hat{\mathbf{x}}_n^{miss})^\top\|_2^2$
- **end for**
- choose $\hat{r} = \arg\min_{r} \sum_{k=1}^K |J_k|^{-1} Err_k(r)$

CV for PCA Repaired



For every fold:

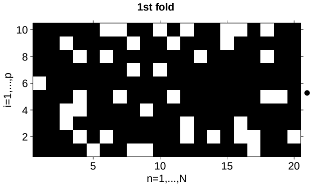
- use **black** entries to obtain $\hat{\mu}$ and $\widehat{\Sigma}$
- predict **white** (missing) entries using **grey** (observed) entries and $\hat{\mu}$ and $\widehat{\Sigma}$ (truncated)
- check the quality of your prediction

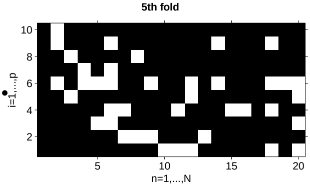
CV for PCA Repaired

```
CV_PCA_repaired <- function(X, Ranks=2:4, K=5){ #X assumed centered
  N   <- nrow(X)
  p   <- ncol(X)
  Ind <- matrix(sample(1:N),nrow=K)
  Err <- array(0,c(K,length(Ranks)))
  for(k in 1:K){
    Xact <- X[-Ind[k,],]
    Xout <- X[Ind[k,],]
    for(r in 1:length(Ranks)){
      C_hat <- sample_cov(Xact)
      EIG   <- eigen(C_hat)
      C_hat <- EIG$vectors[,1:Ranks[r]] %*% diag(EIG$values[1:Ranks[r]]) %*% t(EIG$vectors[,1:Ranks[r]])
      X_hat <- array(0,dim(Xout))
      for(m in 1:dim(Xout)[1]){
        ind <- sample(1:p,floor(p/2)) #partition into observed and missing parts
        Sigma22 <- C_hat[ind,ind]
        Sigma12 <- C_hat[-ind,ind]
        X_hat[m,-ind] <- Sigma12 %*% ginv(Sigma22) %*% Xout[m,ind]
        X_hat[m,ind] <- Xout[m,ind]
      }
      Err[k,r] <- sum((Xout-X_hat)^2)
    }
  }
  return(colSums(Err))
}
```

Improvements?

- Grey entries provide information on μ and Σ, shouldn't we use it?
- Isn't it awkward to first split rows and then columns? Why not just split the bivariate index set?





To cope with this, we need to know how to do **MLE with missing data**

Section 2

Expectation-Maximization (EM) Algorithm

EM Algorithm

Iterative algorithm for calculating Maximum-Likelihood-Estimators (MLEs) in situations, where

- there is **missing data** complicating the calculations (Example 1 and 3 below) or
- it is beneficial to think of our data as if there were some components missing/latent (Example 2 below)
  - when knowing that missing components would render the problem simple

We will assume that solving MLE with the **complete data** is simple

EM will allow us to act as if we knew everything – even when we don't or when we cannot use all the information

Notations

- $\mathbf{X}_{obs}$ are the **observed** random variables
- $\mathbf{X}_{miss}$ are the **missing** random variables
- $\ell_{comp}(\theta)$ is the **complete** log-likelihood of $\mathbf{X} = (\mathbf{X}_{obs}, \mathbf{X}_{miss})$
  - maximizing this to obtain MLE is supposed to be *simple*
  - θ denotes all the parameters, e.g., contains μ and Σ
- $\ell_{obs}(\theta)$ is the **observed** log-likelihood of $\mathbf{X}_{obs}$

We know that

$$\begin{aligned}\ell_{comp}(\theta) &= \ell(\theta \mid \mathbf{X}_{obs}, \mathbf{X}_{miss}) = \ln\{f(\mathbf{X} \mid \theta)\} = \ln\{f(\mathbf{X}_{obs}, \mathbf{X}_{miss}, M \mid \theta, \phi)\} \\ &= \ln\{f(\mathbf{X}_{obs} \mid \theta)\} + \ln\{f(\mathbf{X}_{miss} \mid \mathbf{X}_{obs}, \theta)\} \\ &= \ell_{obs}(\theta) + \ln\{f(\mathbf{X}_{miss} \mid \mathbf{X}_{obs}, \theta)\}\end{aligned}$$

Then, $\ell_{obs}(\theta) = \ell_{comp}(\theta) - \ln\{f(\mathbf{X}_{miss} \mid \mathbf{X}_{obs}, \theta)\}$

Our task is to maximize $\ell_{obs}(\theta)$

Algorithm

Although $\ell_{comp}(\theta)$ is easy to compute, we only observe $\mathbf{X}_{obs}$ and not $\mathbf{X}$

$\Rightarrow$ Let's take on both sides the expectation **given the observed data** and with respect to the **probability measure of $\mathbf{X}$ given by a fixed $\tilde{\theta}$**

Algorithm

Although $\ell_{comp}(\theta)$ is easy to compute, we only observe $\mathbf{X}_{obs}$ and not $\mathbf{X}$

$\Rightarrow$ Let's take on both sides the expectation **given the observed data** and with respect to the **probability measure of $\mathbf{X}$ given by a fixed $\tilde{\theta}$**

EM Algorithm: Start from an initial estimate $\widehat{\theta}^{(0)}$ and for $l = 1, 2, \ldots$ iterate the following two steps until convergence:

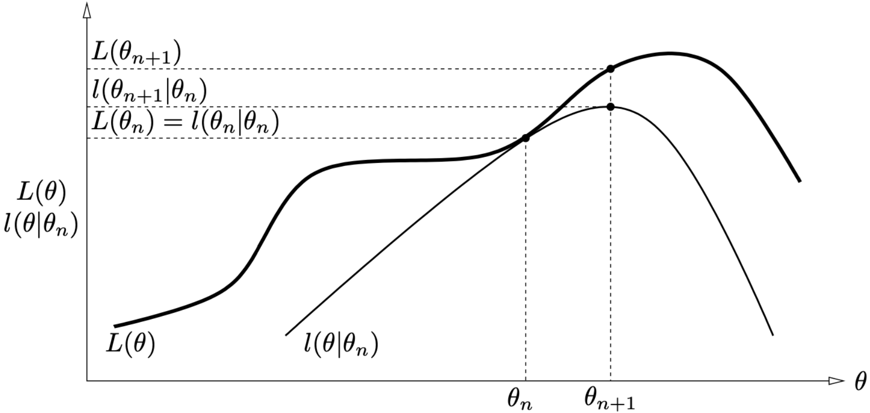
- **E-step**: calculate $\mathbb{E}_{\widehat{\theta}^{(l-1)}}\left[\ell_{comp}(\theta) \middle| \mathbf{X}_{obs} = \mathbf{x}_{obs}\right] =: Q\left(\theta, \widehat{\theta}^{(l-1)}\right)$
- **M-step**: optimize $\arg\max_{\theta} Q\left(\theta, \widehat{\theta}^{(l-1)}\right) =: \widehat{\theta}^{(l)}$

Theorem (Monotone convergence property)

If $\ln\{f(\mathbf{X} \mid \theta)\}$ as well as $\ln\{f(\mathbf{X} \mid \mathbf{X}_{obs}, \theta)\}$ have finite θ'-conditional expectation given $\mathbf{X}_{obs}$ then

$$Q(\theta, \theta') > Q(\theta', \theta') \quad \Rightarrow \quad \ell_{obs}(\theta) > \ell_{obs}(\theta')$$

Graphical interpretation



- $Q(\theta, \theta_n) - H(\theta_n, \theta_n) = \ell(\theta \mid \theta_n) \leq \ell_{obs}(\theta) = L(\theta)$

Ex.1: Censored Observations

Suppose you want to estimate the mean waiting time at an EPFL food truck:

- observed waiting times $\mathbf{x}_{obs} = (x^1_{obs}, \dots, x^{N_{obs}}_{obs})^\top$ for $\mathbf{X}_{obs}$
- food truck closes when N_{miss} individuals are still queuing, such that $\mathbf{X}_{miss} = (X^1_{miss}, \dots, X^{N_{miss}}_{miss})^\top$ are not observed but only a vector of right-censored waiting times $\tilde{\mathbf{x}}_{miss}$ with $\forall n : X^{(n)}_{miss} > \tilde{x}^{(n)}_{miss}$
- overall $N = N_{obs} + N_{miss}$ individuals considered (known)

$\Rightarrow$ Apply EM-algorithm assuming waiting times are i.i.d. and follow an exponential distribution with density $f(x) = \lambda \exp(-\lambda x)$

Ex.1: Censored Observations – E-step

- **E-step**: calculate
  $\mathbb{E}_{\widehat{\lambda}^{(l-1)}}\left[\ell_{comp}(\lambda)\middle|\mathbf{X}_{obs}=\mathbf{x}_{obs}, \forall n : X_{miss}^{(n)} > \tilde{x}_{miss}^{(n)}\right] =: Q(\lambda, \hat{\lambda}^{(l-1)})$

For iterations $l = 1, 2, ...$

$$
\begin{aligned}
Q(\lambda, \hat{\lambda}^{(l-1)}) &= \mathbb{E}_{\widehat{\lambda}^{(l-1)}}\left[\ell_{comp}(\lambda) \mid \mathbf{x}_{obs}, \tilde{\mathbf{x}}_{miss}\right] \\
&= \mathbb{E}_{\widehat{\lambda}^{(l-1)}}\Big[\underbrace{N\log(\lambda) - \lambda\sum_{n=1}^{N_{obs}} X_{obs}^{(n)} - \lambda\sum_{n=1}^{N_{miss}} X_{miss}^{(n)}}_{\log\{\prod_{n=1}^{N_{obs}} f(X_{obs}^{(n)})\cdot\prod_{n=1}^{N_{miss}} f(X_{miss}^{(n)})\}} \mid \mathbf{x}_{obs}, \tilde{\mathbf{x}}_{miss}\Big] \\
&= N\log(\lambda) - \lambda\sum_{n=1}^{N_{obs}} x_{obs}^{(n)} - \lambda\sum_{n=1}^{N_{miss}} \underbrace{\mathbb{E}_{\widehat{\lambda}^{(l-1)}}\left[X_{miss}^{(n)} \mid \tilde{\mathbf{x}}_{miss}\right]}_{\overset{X\sim Exp(\hat{\lambda}^{(l-1)})}{\underset{\text{"}memoryless\text{"}}{=}} 1/\hat{\lambda}^{(l-1)} + \tilde{x}_{miss}^{(n)}} \\
&= N\log(\lambda) - \lambda\left(N_{obs}\bar{x}_{obs} + N_{miss}\frac{1}{\hat{\lambda}^{(l-1)}} + N_{miss}\bar{\tilde{x}}_{miss}\right)
\end{aligned}
$$

Ex.1: Censored observations – M-step

- **M-step**: optimize $\arg\max_{\lambda} \; Q(\lambda, \hat{\lambda}^{(l-1)})$

$$Q(\lambda, \hat{\lambda}^{(l-1)}) = N \log(\lambda) - \lambda\big(N_{obs}\bar{x}_{obs} + \frac{N_{miss}}{\hat{\lambda}^{(l-1)}} + N_{miss}\bar{\tilde{x}}_{miss}\big)$$

$$\Rightarrow \quad \frac{\partial Q}{\partial \lambda}(\lambda, \hat{\lambda}^{(l-1)}) = \frac{N}{\lambda} - (N_{obs}\bar{x}_{obs} + N_{miss}\frac{1}{\hat{\lambda}^{(l-1)}} + N_{miss}\bar{\tilde{x}}_{miss}) \stackrel{!}{=} 0$$

$$\Rightarrow \quad \hat{\lambda}^{(l)} = \frac{N}{N_{obs}\bar{x}_{obs} + \frac{N_{miss}}{\hat{\lambda}^{(l-1)}} + N_{miss}\bar{\tilde{x}}_{miss}}$$

We can compute the stationary point $\hat{\lambda}^{(l)} = \hat{\lambda}^{(l-1)} = \hat{\lambda}$

$$\hat{\lambda} = \frac{N_{obs}}{N_{obs}\bar{x}_{obs} + N_{miss}\bar{\tilde{x}}_{miss}}$$

which could also be obtained by maximizing the ML function with censored data!

Ex.2: Mixture distributions

One of the most popular applications of the EM-algorithm:

Estimating mixture distributions for modelling multimodality or clustering/classification (soft or hard)

Mixture of two Gaussian distributions:

Let $X^{(1)}, \dots, X^{(N)}$ be i.i.d. random variables each with pdf

$$f_\theta(x) = (1 - \tau)\ \varphi_{\mu_1,\sigma_1}(x) + \tau\ \varphi_{\mu_2,\sigma_2}(x)$$

where $\theta = (\tau, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)^\top$, with

- $\varphi_{\mu,\sigma}$ is the pdf of a Gaussian with mean μ and standard deviation σ,
- μ_1, μ_2 and σ_1^2, σ_2^2 are the means and variances of the mixture components, and
- $\tau \in (0, 1)$ is the mixing proportion

Note: case of mixture of m Gaussians is easily generalizable, though M-step is trickier

Ex.2: Mixture distributions – factorization via latent variables

Log-likelihood has no nice form:

$$\ell_{obs}(\theta) = \sum_{n=1}^{N} \log \left\{ (1-\tau)\, \varphi_{\mu_1,\sigma_1}\left(X^{(n)}\right) + \tau\, \varphi_{\mu_2,\sigma_2}\left(X^{(n)}\right) \right\}$$

Trick: add latent i.i.d. indicators $Z^{(n)} \sim Bernoulli(\tau)$ such that $X^{(n)} \mid Z^{(n)} = 0 \sim N(\mu_1, \sigma_1^2)$ and $X^{(n)} \mid Z^{(n)} = 1 \sim N(\mu_2, \sigma_2^2)$

Given $Z^{(n)} = z^{(n)}$, $n = 1, \dots, N$, the joint likelihood can be written as

$$L_{comp}(\theta) = (1-\tau)^{N_1} \tau^{N_2} \prod_{n=1}^{N} \varphi_{\mu_1,\sigma_1}\left\{X^{(n)}\right\}^{(1-Z^{(n)})} \varphi_{\mu_2,\sigma_2}\left\{X^{(n)}\right\}^{Z^{(n)}}$$

with $N_2 = \sum_{n=1}^{N} Z^{(n)}$ and $N_1 = N - N_2$

Ex.2: Mixture distributions – E-step – Part I

- **E-step**: calculate $\mathbb{E}_{\hat{\theta}^{(l-1)}}\left[\ell_{comp}(\theta) \middle| \mathbf{X} = \mathbf{x}\right] =: Q(\theta, \widehat{\theta}^{(l-1)})$

$$\ell_{comp}(\theta) = \ln L_{comp}(\theta) = N_1 \ln(1-\tau) + N_2 \ln(\tau) + \\ + \sum_{n=1}^{N} (1 - Z^{(n)}) \ln \varphi_{\mu_1,\sigma_1}\left(X^{(n)}\right) + \sum_{n=1}^{N} Z^{(n)} \ln \varphi_{\mu_2,\sigma_2}\left(X^{(n)}\right)$$

such that, we obtain

$$\mathbb{E}_{\hat{\theta}^{(l-1)}}\left[\ell_{comp}(\theta) \middle| \mathbf{X} = \mathbf{x}\right] = \log(1-\tau)(N - \sum_{n=1}^{N} p_n^{(l-1)}) + \log(\tau) \sum_{n=1}^{N} p_n^{(l-1)} + \\ + \sum_{n=1}^{N} (1 - p_n^{(l-1)}) \log \varphi_{\mu_1,\sigma_1}\left(x^{(n)}\right) + \sum_{n=1}^{N} p_n^{(l-1)} \log \varphi_{\mu_2,\sigma_2}\left(x^{(n)}\right)$$

with $p_n^{(l-1)} = \mathbb{E}_{\hat{\theta}^{(l-1)}}\left[Z^{(n)} \middle| X^{(n)} = x^{(n)}\right] \overset{Bayes}{=} \frac{\varphi_{\hat{\mu}_2^{(l-1)},\hat{\sigma}_2^{(l-1)}}\left(x^{(n)}\right)\hat{\tau}^{(l-1)}}{f_{\hat{\theta}^{(l-1)}}(x^{(n)})}$.

Ex.2: Mixture distributions – M-step

- **M-step**: optimize $\arg\max_{\theta} Q(\theta, \hat{\theta}^{(l-1)})$

Hence, $Q(\theta, \hat{\theta}^{(l-1)})$ nicely splits into three parts

$$
\begin{aligned}
&Q(\theta, \hat{\theta}^{(l-1)}) = \\
&\quad \mathbf{A}: \quad \log(1-\tau)(N - \sum_{n=1}^{N} p_n^{(l-1)}) + \log(\tau) \sum_{n=1}^{N} p_n^{(l-1)} + \\
&\quad \mathbf{B}: \quad \sum_{n=1}^{N} (1 - p_n^{(l-1)}) \log \varphi_{\mu_1, \sigma_1} \left\{ x^{(n)} \right\} + \\
&\quad \mathbf{C}: \quad \sum_{n=1}^{N} p_n^{(l-1)} \log \varphi_{\mu_2, \sigma_2} \left\{ x^{(n)} \right\}
\end{aligned}
$$

which can be optimized separately, where $\mathbf{A}$ has the form of a binomial and $\mathbf{B}$ and $\mathbf{C}$ of (weighted) Gaussian log-likelihood $\Rightarrow$ optimize accordingly

Ex.3: Multivariate Gaussian with Missing Entries

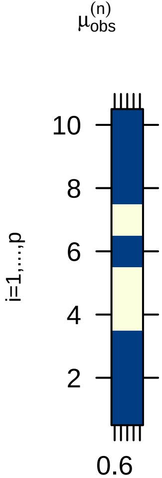
Let $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(N)}$ be i.i.d. p-variate normally distributed with mean μ and covariance Σ

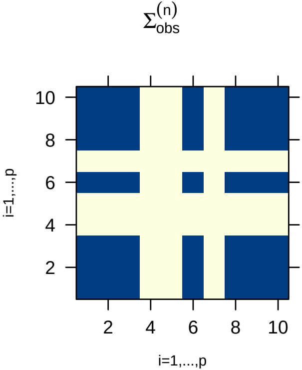
For each n, only a realization $\mathbf{x}_{obs}^{(n)}$ of $\mathbf{X}_{obs}^{(n)}$, subvector of $\mathbf{X}^{(n)}$, is observed

The goal is to estimate μ and Σ from the incomplete observations

Ex.3: Multivariate Gaussian with Missing Entries

Let $\mu_{obs}^{(n)}$ and $\Sigma_{obs}^{(n)}$ denote the mean and covariance of $\mathbf{X}_{obs}^{(n)}$, i.e., $\mu_{obs}^{(n)}$ is just a sub-vector of μ and $\Sigma_{obs}^{(n)}$ is a sub-matrix of Σ





Ex.3: Multivariate Gaussian with Missing Entries

Recall the density $f(\mathbf{x})$ of a p-variate Gaussian:

$$f(\mathbf{x}^{(n)}) \propto \det(\Sigma)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\mathbf{x}^{(n)} - \mu\right)^{\top}\Sigma^{-1}\left(\mathbf{x}^{(n)} - \mu\right)\right\},$$

Hence, log-likelihoods are given by

$$\ell_{obs}(\mu, \Sigma) = \text{const } - \frac{1}{2}\sum_{n=1}^{N} \log\det(\Sigma_{obs}^{(n)}) -$$

$$- \sum_{n=1}^{N} \frac{1}{2}\left(\mathbf{x}_{obs}^{(n)} - \mu_{obs}^{(n)}\right)^{\top}\left(\Sigma_{obs}^{(n)}\right)^{-1}\left(\mathbf{x}_{obs}^{(n)} - \mu_{obs}^{(n)}\right)$$

$$\ell_{comp}(\mu, \Sigma) = \text{const } - \frac{N}{2}\ln\,\det(\Sigma) - \sum_{n=1}^{N}\frac{1}{2}\underbrace{\left(\mathbf{x}^{(n)} - \mu\right)^{\top}\Sigma^{-1}\left(\mathbf{x}^{(n)} - \mu\right)}_{\operatorname{tr}\left\{\left(\mathbf{x}^{(n)}-\mu\right)\left(\mathbf{x}^{(n)}-\mu\right)^{\top}\Sigma^{-1}\right\}}.$$

Optimizing ℓ_{comp} is easier than optimizing ℓ_{obs} $\Rightarrow$ EM-Algorithm

Ex.3: Multivariate Gaussian with Missing Entries – E-step

- **E-step**: calculate $\mathbb{E}_{\hat{\theta}^{(l-1)}}\{\ell_{comp}(\theta) \big| \forall n : \mathbf{X}_{obs}^{(n)} = \mathbf{x}_{obs}^{(n)}\} =: Q(\theta, \hat{\theta}^{(l-1)})$ with $\theta = (\mu, \Sigma)^\top$

$$Q(\theta, \hat{\theta}^{(l-1)}) = \text{const} - \frac{N}{2} \ln \det(\Sigma)$$

$$- \sum_{n=1}^{N} \frac{1}{2} \text{tr}\Big[\underbrace{\mathbb{E}_{\hat{\theta}^{(l-1)}}\Big\{ \big(\mathbf{X}^{(n)} - \mu\big)\big(\mathbf{X}^{(n)} - \mu\big)^\top \Big| \forall n : \mathbf{X}_{obs}^{(n)} = \mathbf{x}_{obs}^{(n)} \Big\}}_{\overset{\text{some calculations}}{=} (\hat{\mathbf{x}}^{(n)(l-1)} - \mu)(\hat{\mathbf{x}}^{(n)(l-1)} - \mu)^\top + \mathbf{C}^{(n)}} \Sigma^{-1} \Big]$$

with $\hat{\mathbf{x}}^{(n)(l-1)} = \mathbb{E}_{\hat{\theta}^{(l-1)}}\big(\mathbf{X}^{(n)} \big| \forall n : \mathbf{X}_{obs}^{(n)} = \mathbf{x}_{obs}^{(n)}\big)$ and

$$\mathbf{C}^{(n)} = \Big\{ \text{Cov}_{\hat{\theta}^{(l-1)}} \Big(X_i^{(n)}, X_j^{(n)} \mid \forall n : \mathbf{X}_{obs}^{(n)} = \mathbf{x}_{obs}^{(n)} \Big) \Big\}_{i,j}$$

Ex.3: Multivariate Gaussian with Missing Entries – M-step

- **M-step**: optimize $\arg\max_{\theta} Q(\theta, \hat{\theta}^{(l-1)})$

$$
\begin{aligned}
Q(\theta,\hat{\theta}^{(l-1)}) = \text{const } - \frac{N}{2}\log\det(\Sigma) - \\
- \sum_{n=1}^{N} \frac{1}{2}\mathrm{tr}\left[\left\{(\hat{\mathbf{x}}^{(n)(l-1)} - \mu)(\hat{\mathbf{x}}^{(n)(l-1)} - \mu)^{\top} + \mathbf{C}^{(n)}\right\}\Sigma^{-1}\right]
\end{aligned}
$$

has a similar form as a multivariate normal and estimators can be derived accordingly, resulting in

$$
\hat{\mu}^{(l)} = N^{-1}\sum_{n=1}^{N}\hat{\mathbf{x}}^{(n)(l-1)}
$$

and

$$
\widehat{\Sigma}^{(l)} = \frac{1}{N}\sum_{n=1}^{N}\left\{(\hat{\mathbf{x}}^{(n)(l-1)} - \hat{\mu}^{(l)})(\hat{\mathbf{x}}^{(n)(l-1)} - \hat{\mu}^{(l)})^{\top} + \mathbf{C}^{(n)}\right\}
$$

Recap

Example 1:

- part of data missing but their censored versions carry some information
- the likelihood is linear (w.r.t. observations) and thus the **E-step** coincides with imputation (missing data replaced by their expectations)
  - this is rare! It works when the log-likelihood is linear in the missing data

Example 2:

- there is no true missing data here, but it is beneficial to imagine it
- the likelihood is linear w.r.t. the imagined observations $\Rightarrow$ simplification

Example 3:

- likelihood of observed data easy to formulate, yet hard to optimize directly
- no linearity in log-likelihood $\Rightarrow$ no imputation, more effort to compute expected likelihood (though still relatively simple, since exponential family)

References

- Dempster, A. P., N. M. Laird & D. B. Rubin. (1977) "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1: 1-22
  - one of the most cited papers in statistics of all times
- Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data*. 3rd Edition
- McLachlan, G. J. & Krishnan, T. (2008) *The EM Algorithm and Extensions*. 2nd Edition

Exercise: Multinomial distribution

Go to Exercise 3 for details.

Assignment 5 [5 %]

Go to Assignment 5 for details.