

# Week 6: The EM-Algorithm

MATH-517 Statistical Computation and Visualization

Linda Mhalla

2023-10-27

## Section 1

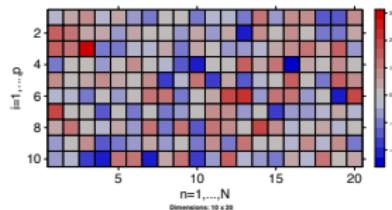
Motivation From Last Week

# CV for PCA Repaired

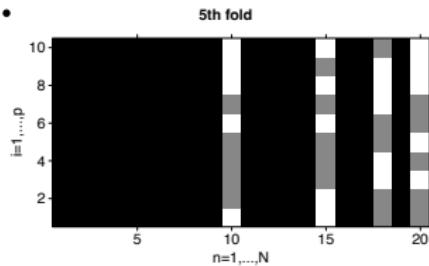
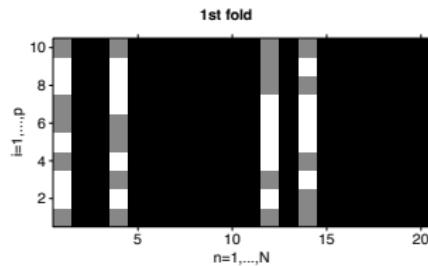
Assume that data  $\mathbf{x}_n \in \mathbb{R}^p$  are i.i.d. realizations of  $X \sim \mathcal{N}(\mu, \Sigma)$

- split data into  $K$  folds:  $J_1, \dots, J_K$
- **for**  $k = 1, \dots, K$ 
  - estimate  $\mu$  and  $\Sigma$  empirically using all but the  $k$ -th fold  $J_k$ , and truncate  $\Sigma$  to be rank- $r$
  - **for**  $n \in J_k$ 
    - split  $\mathbf{x}_n$  into a “missing” part  $\mathbf{x}_n^{miss}$  that will be used for validation and an “observed” part  $\mathbf{x}_n^{obs}$
    - 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



Use white points (observed) to predict grey points (missing) - predict mu and sigma



For every fold:

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

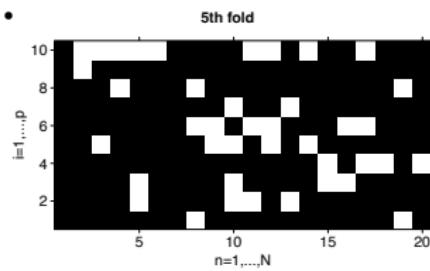
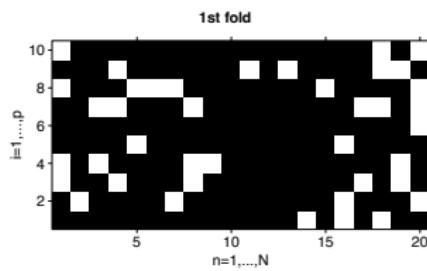
# CV for PCA Repaired

```
CV_PCA_repaired <- function(X, Ranks=2:4, K=5){  
  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],]  
    SVD <- svd(Xact)  
    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))  
        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  $\mu$  and  $\Sigma$ , shouldn't we use it?
- Isn't it awkward to first split rows and then columns? Why not just split the bivariate index set?

white = missing points



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*
  - $\theta$  denotes all the parameters, e.g., contains  $\mu$  and  $\Sigma$
- $\ell_{obs}(\theta)$  is the **observed** log-likelihood of  $\mathbf{X}_{obs}$

We know that

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

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  $\mathbf{X}$

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

# Algorithm

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

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

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

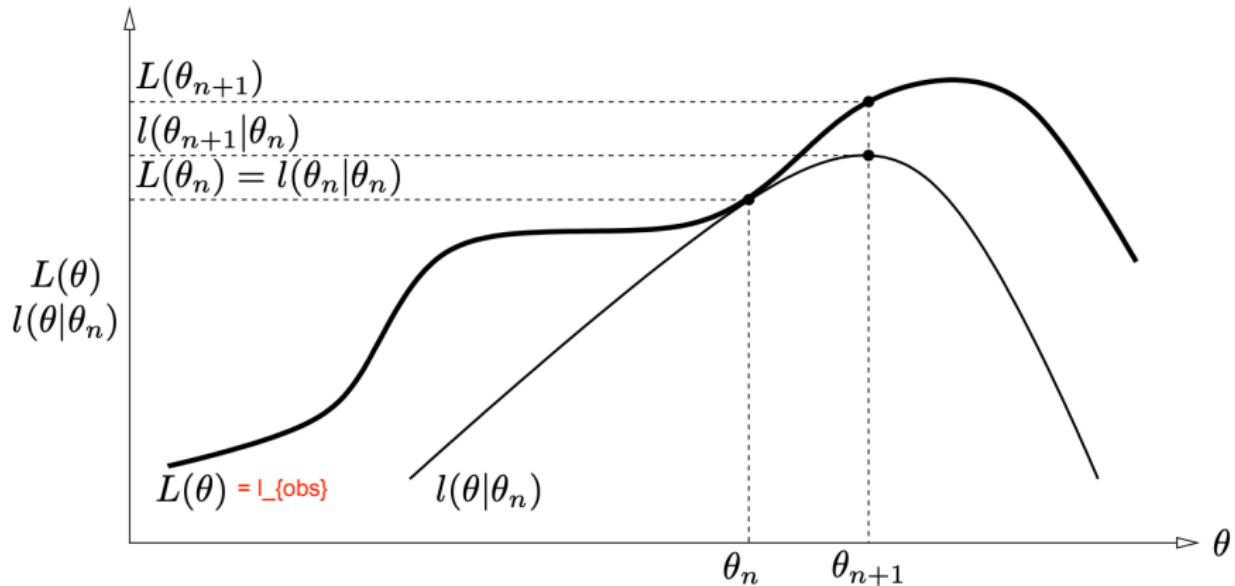
- **E-step:** calculate  $\mathbb{E}_{\hat{\theta}^{(l-1)}}[\ell_{comp}(\theta) | \mathbf{X}_{obs} = \mathbf{x}_{obs}] =: Q(\theta, \hat{\theta}^{(l-1)})$
- **M-step:** optimize  $\arg \max_{\theta} Q(\theta, \hat{\theta}^{(l-1)}) =: \hat{\theta}^{(l)}$

## Theorem (Monotone convergence property)

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

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

# Graphical interpretation



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

Q is always be below than L (the one we want to maximize)

Q and L will only meet at theta\_n

## Ex.1: Censored Observations

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

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

⇒ 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

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

For iterations  $l = 1, 2, \dots$

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

Because they are i.i.d,  
we don't need  $\mathbf{x}_{obs}$

$$\begin{aligned}
 & E\text{-step: } E_{\lambda^{(t-1)}} \left\{ p_{\text{comp}}(\lambda) \mid X_{\text{obs}} = x_{\text{obs}}, \forall n \quad X_{\text{miss}}^{(n)} > \hat{x}_{\text{miss}}^{(n)} \right\} := Q(\lambda, \lambda^{(t)}) \\
 & = E_{\lambda^{(t-1)}} \left\{ \sum_{i=1}^N \log(\lambda) - \lambda x^{(i)} \mid X_{\text{obs}} = x_{\text{obs}}, \forall n \quad X_{\text{miss}}^{(n)} > \hat{x}_{\text{miss}}^{(n)} \right\} \\
 & = E_{\lambda^{(t-1)}} \left\{ N \log(\lambda) - \lambda \sum_{n=1}^{N_{\text{obs}}} x_{\text{obs}}^{(n)} - \lambda \sum_{n=1}^{N_{\text{miss}}} \hat{x}_{\text{miss}}^{(n)} \mid X_{\text{obs}} = x_{\text{obs}}, \forall n \quad X_{\text{miss}}^{(n)} > \hat{x}_{\text{miss}}^{(n)} \right\} \\
 & = N \log(\lambda) - \lambda \sum_{n=1}^{N_{\text{obs}}} x_{\text{obs}}^{(n)} - \lambda \sum_{n=1}^{N_{\text{miss}}} \underbrace{E_{\lambda^{(t-1)}}(X_{\text{miss}}^{(n)} \mid X_{\text{miss}}^{(n)} > \hat{x}_{\text{miss}}^{(n)})}_{\text{yellow box}}
 \end{aligned}$$

$$\begin{aligned}
 \textcircled{2} \quad & \Pr(X_{\text{miss}}^{(j)} > t \mid X_{\text{miss}}^{(j)} > \hat{x}_{\text{miss}}^{(j)}) \\
 & = \Pr(X_{\text{miss}}^{(j)} > (t - \hat{x}_{\text{miss}}^{(j)}) + \hat{x}_{\text{miss}}^{(j)} \mid X_{\text{miss}}^{(j)} > \hat{x}_{\text{miss}}^{(j)}) \\
 & = \Pr(X_{\text{miss}}^{(j)} > t - \hat{x}_{\text{miss}}^{(j)}) \approx \Pr(X_{\text{miss}}^{(j)} + \hat{x}_{\text{miss}}^{(j)} > t)
 \end{aligned}$$

$$\Rightarrow E_{\lambda^{(t-1)}}(X_{\text{miss}}^{(n)} \mid X_{\text{miss}}^{(n)} > \hat{x}_{\text{miss}}^{(n)}) = \hat{x}_{\text{miss}}^{(n)} + \frac{1}{\lambda^{(t-1)}}$$

## Ex.1: Censored observations – M-step

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

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

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

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

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  $\mu$  and standard deviation  $\sigma$ ,
- $\mu_1, \mu_2$  and  $\sigma_1^2, \sigma_2^2$  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:

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

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

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

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

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

$$E_{\hat{\theta}^{(c-1)}} \{ \ell_{\text{comp}}(\theta) | X=x \} = -$$

$$= E_{\hat{\theta}^{(c-1)}} \left\{ N_2 \log(\varepsilon) + N_1 \log(1-\varepsilon) + \sum_{n=1}^N (1-z^{(n)}) \ln \frac{1}{N_n \varepsilon_n} \right.$$

$$\left. + z^{(n)} \ln \frac{1}{N_n (1-\varepsilon_n)} | X=x \right\}$$

If all boils down to

$$E_{\hat{\theta}^{(c-1)}} (z^{(n)} | X^{(n)}, \hat{\theta}^{(c-1)}) = \Pr(z^{(n)}=1 | X^{(n)}, \hat{\theta}^{(c-1)})$$

$$\stackrel{\text{Bayes}}{=} \frac{f_{X|Z=1}(x^{(n)} | \hat{\theta}^{(c-1)}) \Pr(z^{(n)}=1 | \hat{\theta}^{(c-1)})}{f_{\hat{\theta}^{(c-1)}}(x^{(n)})}$$

$$= \frac{f_{N_2 \mid \hat{\theta}^{(c-1)}}(x^{(n)}) \hat{\varepsilon}^{(c-1)}}{f_{\hat{\theta}^{(c-1)}}(x^{(n)})}$$

$$:= p_n^{(c-1)}$$

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

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

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

such that, we obtain

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

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

## 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

$$Q(\theta, \hat{\theta}^{(l-1)}) =$$

Only depends on tau    **A** :  $\log(1 - \tau)(N - \sum_{n=1}^N p_n^{(l-1)}) + \log(\tau) \sum_{n=1}^N p_n^{(l-1)} +$

**B** :  $\sum_{n=1}^N (1 - p_n^{(l-1)}) \log \varphi_{\mu_1, \sigma_1} \{x^{(n)}\} +$

**C** :  $\sum_{n=1}^N p_n^{(l-1)} \log \varphi_{\mu_2, \sigma_2} \{x^{(n)}\}$

Weighted  
gaussian

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

## Ex.3: Multivariate Gaussian with Missing Entries

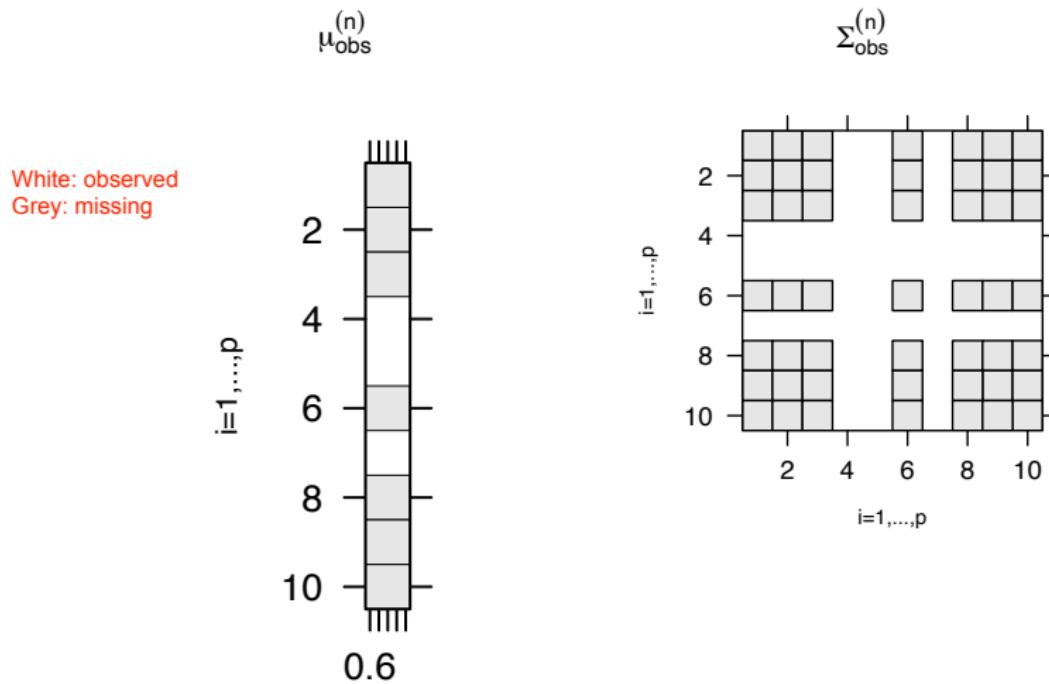
Let  $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(N)}$  be i.i.d.  $p$ -variate normally distributed with mean  $\mu$  and covariance  $\Sigma$

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  $\mu$  and  $\Sigma$  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  $\mu$  and  $\Sigma_{obs}^{(n)}$  is a sub-matrix of  $\Sigma$



## Ex.3: Multivariate Gaussian with Missing Entries

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

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

Hence, log-likelihoods are given by

$$\begin{aligned}\ell_{obs}(\mu, \Sigma) &= \text{const} - \frac{1}{2} \sum_{n=1}^N \log \det(\Sigma_{obs}^{(n)}) - \\ &\quad - \sum_{n=1}^N \frac{1}{2} (\mathbf{x}_{obs}^{(n)} - \boldsymbol{\mu}_{obs}^{(n)})^\top (\Sigma_{obs}^{(n)})^{-1} (\mathbf{x}_{obs}^{(n)} - \boldsymbol{\mu}_{obs}^{(n)}) \\ \ell_{comp}(\mu, \Sigma) &= \text{const} - \frac{N}{2} \ln \det(\Sigma) - \sum_{n=1}^N \frac{1}{2} \underbrace{(\mathbf{x}^{(n)} - \mu)^\top \Sigma^{-1} (\mathbf{x}^{(n)} - \mu)}_{\text{tr} \left\{ (\mathbf{x}^{(n)} - \mu) (\mathbf{x}^{(n)} - \mu)^\top \Sigma^{-1} \right\}}.\end{aligned}$$

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

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

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

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

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

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

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

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

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

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

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

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

# 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 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.