Week 7: The EM-Algorithm

MATH-517 Statistical Computation and Visualization

Tomas Masak

November 4th 2022

Section 1

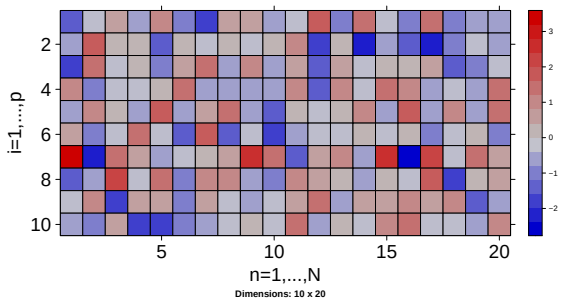
Motivation From the Last Week

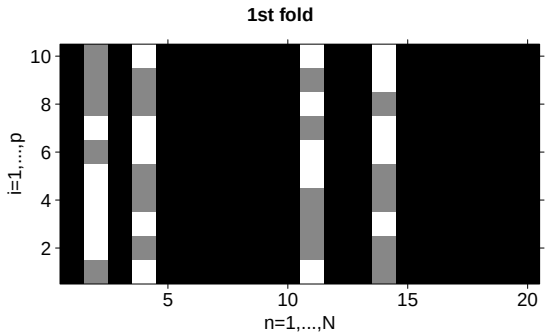
CV for PCA Repaired

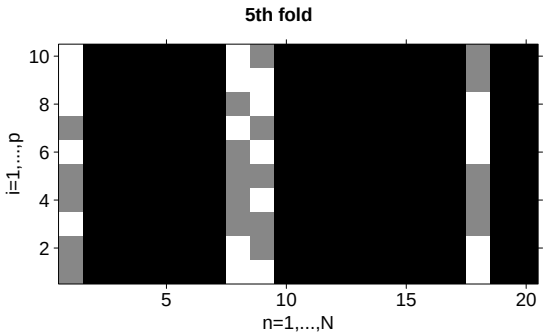
Assume that data $\mathbf{x}_n \in \mathbb{R}^p$ are i.i.d. realizations of $X \sim \mathcal{N}(\mu, \Sigma)$.

- split data into K folds $J_1, \ldots, J_K$
- **for** $k = 1, \ldots, K$
  - estimate μ and Σ empirically using all but the k-th fold J_k, but truncate Σ to be rank-r
  - **for** $n \in J_k$
    - split $\mathbf{x}_n$ 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 $Err_k(r) = \sum_{n \in J_k} \|\mathbf{x}_n^{obs} - \widehat{\mathbf{x}}_n^{obs}\|_2^2$
- **end for**
- choose $\widehat{r} = \arg\min_r \sum_{k=1}^K |J_k|^{-1} Err_k(r)$

CV for PCA Repaired





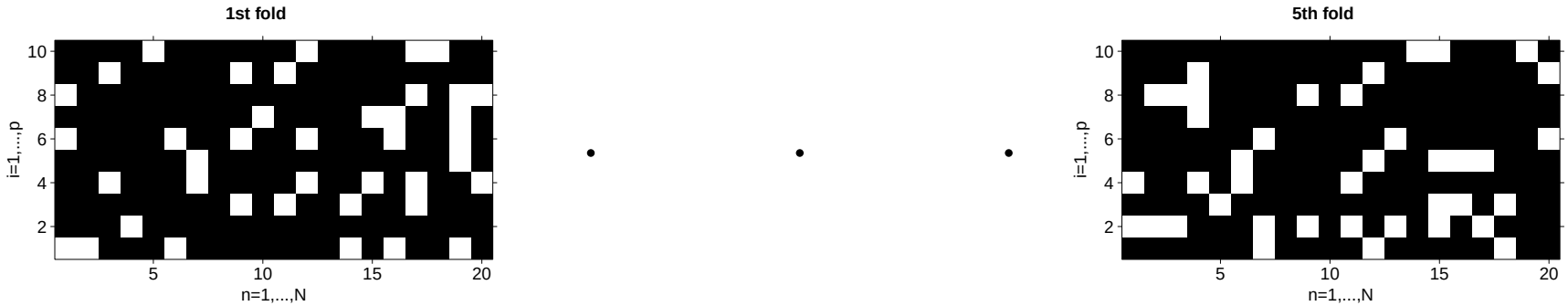


For every fold:

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

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 to 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 (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 like if we knew everything – even when we don't or when we cannot use all the information.

Notation and the Algorithm

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

Our task is to maximize $\ell_{obs}(\theta)$, the **observed** log-likelihood of $\boldsymbol{X}_{obs}$.

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

Ex.1: Censored Observations

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

- observed waiting times $\mathbf{x}_{obs} = (x_{obs}^1, \ldots, x_{obs}^{N_{obs}})^\top$ for $\boldsymbol{X}_{obs}$.
- food truck closes when N_{miss} individuals are still queuing, such that $\boldsymbol{X}_{miss} = (X_{miss}^1, \ldots, X_{miss}^{N_{miss}})^\top$ 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.

$\Rightarrow$ Apply EM-algorithm assuming weighting times iid. distributed follwing an exponential distribution with density $f(x) = \lambda \exp(-\lambda x)$.

Ex.1: Censored Observations – E-step

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

For iterations $l=1,\ldots$:

$$\mathbb{E}_{\widehat{\lambda}^{(l-1)}}\left[\ell_{comp}(\theta)\mid\mathbf{x}_{obs},\tilde{\mathbf{x}}_{miss}\right]=$$

$$=\mathbb{E}_{\widehat{\lambda}^{(l-1)}}\Big[\underbrace{N\log(\lambda)-\lambda\sum_{n=1}^{N_{obs}}X_{obs}^{(n)}-\lambda\sum_{n=1}^{N_{miss}}X_{miss}^{(n)}}_{\log\left(\prod_{n=1}^{N_{obs}}f(X_{obs}^{(n)})\cdot\prod_{n=1}^{N_{miss}}f(X_{miss}^{(n)})\right)}\mid\mathbf{x}_{obs},\tilde{\mathbf{x}}_{miss}\Big]$$

$$=N\log(\lambda)-\lambda\sum_{n=1}^{N_{obs}}x_{obs}^{(n)}-\lambda\sum_{n=1}^{N_{miss}}\underbrace{\mathbb{E}_{\widehat{\lambda}^{(l-1)}}\left[X_{miss}^{(n)}\mid\tilde{\mathbf{x}}_{miss}\right]}_{\substack{X\sim Exponential(\hat{\lambda}^{(l-1)})\\=\\"memoryless"}}1/\hat{\lambda}^{(l-1)}+\tilde{x}_{miss}^{(n)}$$

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

Ex.1: Censored observations – M-step

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

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

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

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

Ex.2: Mixture distributions

One of the most popular applications of the EM-algorithm:
Estimating mixture distributions for modelling multimodality

Mixture of two Gaussian distributions:

Let $X^{(1)}, \ldots, X^{(N)}$ be iid. distributed as X with probability density

$$f(x) = (1-\tau)\,\varphi\left(\frac{x-\mu_1}{\sigma_1}\right) + \tau\,\varphi\left(\frac{x-\mu_2}{\sigma_2}\right)$$

where

- φ is the density of a standard normal, and
- $\mu_1 < \mu_2$ and σ_1^2, σ_2^2 are the means and variances of the mixture components, and
- $\tau \in (0,1)$ is the share of the second component stacked in a vector $\boldsymbol{\theta} = (\tau, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)^\top$

Ex.2: Mixture distributions – factorization via latent variables

However, log-likelihood has no nice form:

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

Trick: add latent iid. component 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 $Z^{(n)} = z^{(n)}$, $n = 1, \ldots, N$, the joint likelihood can be written as

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

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 $\mathbb{E}_{\widehat{\theta}^{(l-1)}}\left[\tilde{\ell}(\theta)\middle|\boldsymbol{X}=\mathbf{x}\right] =: Q\left(\theta, \widehat{\theta}^{(l-1)}\right)$

$$\ell_{comp}(\boldsymbol{\theta}) = \log L_{comp}(\boldsymbol{\theta}) = N_1 \log(1-\tau) + N_2 \log(\tau) + \\ + \sum_{n=1}^{N}(1-Z^{(n)}) \log \varphi\left(\frac{X^{(n)}-\mu_1}{\sigma_1}\right) + \sum_{n=1}^{N} Z^{(m)} \log \varphi\left(\frac{X^{(n)}-\mu_2}{\sigma_2}\right)$$

such that, we obtain

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

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

Ex.2: Mixture distributions – M-step

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

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

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

$$\mathbf{A}: \quad \log(1-\tau)(N - \sum_{n=1}^{N} p_n^{(l-1)}) + \log(\tau) \sum_{n=1}^{N} p_n^{(l-1)} +$$

$$\mathbf{B}: \quad + \sum_{n=1}^{N} (1 - p_n^{(l-1)}) \log \varphi \left(\frac{x^{(n)} - \mu_1}{\sigma_1} \right) +$$

$$\mathbf{C}: \quad + \sum_{n=1}^{N} p_n^{(l-1)} \log \varphi \left(\frac{x^{(n)} - \mu_2}{\sigma_2} \right)$$

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

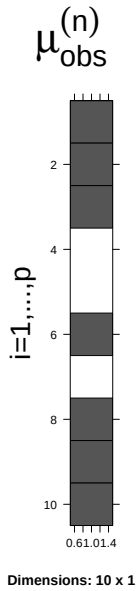
Let $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(N)}$ iid. p-variate normally distributed with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.

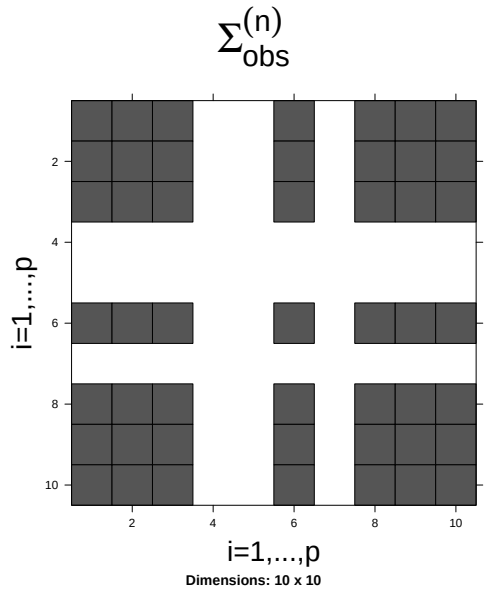
For each n, only a realization $\mathbf{x}_{obs}^{(n)}$ of a subvector $\boldsymbol{X}_{obs}^{(n)}$ of $\boldsymbol{X}^{(n)}$ is observed.

The goal is to estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from the incomplete measurements.

Ex.3: Multivariate Gaussian with Missing Entries

Let further denote $\boldsymbol{\mu}_{obs}^{(n)}$ and $\boldsymbol{\Sigma}_{obs}^{(n)}$ the mean and the covariance of $\boldsymbol{X}_{obs}^{(n)}$, i.e. $\boldsymbol{\mu}_{obs}^{(n)}$ is just a sub-vector of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_{obs}^{(n)}$ is a sub-matrix of $\boldsymbol{\Sigma}$.





Ex.3: Multivariate Gaussian with Missing Entries

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

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

and, hence, log-likelihood are given by

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

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

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

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

- **E-step**: calculate $\mathbb{E}_{\widehat{\boldsymbol{\theta}}^{(l-1)}}\left[\ell_{comp}(\boldsymbol{\theta}) \middle| \forall n : \boldsymbol{X}_{obs}^{(n)} = \mathbf{x}_{obs}^{(n)}\right] =: Q\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{(l-1)}\right)$ with $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

$$Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(l-1)}) = \text{const} - \frac{N}{2}\log\det(\boldsymbol{\Sigma}) -$$
$$- \sum_{n=1}^{N} \frac{1}{2}\text{tr}\Big(\underbrace{\mathbb{E}_{\boldsymbol{\theta}^{(l-1)}}\left[\left(\boldsymbol{X}^{(n)} - \boldsymbol{\mu}\right)\left(\boldsymbol{X}^{(n)} - \boldsymbol{\mu}\right)^{\top} \middle| \forall n : \boldsymbol{X}_{obs}^{(n)} = \mathbf{x}_{obs}^{(n)} \right]}_{\overset{\text{some calculation}}{=} (\widehat{\mathbf{x}}^{(n)(l-1)} - \boldsymbol{\mu})(\widehat{\mathbf{x}}^{(n)(l-1)} - \boldsymbol{\mu})^{\top} + \mathbf{C}^{(n)}} \boldsymbol{\Sigma}^{-1} \Big)$$

with $\hat{\mathbf{x}}^{(n)(l-1)} = \mathbb{E}_{\widehat{\boldsymbol{\theta}}^{(l-1)}}\left[\boldsymbol{X}^{(n)} \middle| \forall n : \boldsymbol{X}_{obs}^{(n)} = \mathbf{x}_{obs}^{(n)}\right]$ and
$\mathbf{C}^{(n)} = \left\{ \text{Cov}_{\widehat{\boldsymbol{\theta}}^{(l-1)}}\left(X_i^{(n)}, X_j^{(n)} \mid \boldsymbol{X}_{obs}^{(n)} = \mathbf{x}_{obs}^{(n)} \right) \right\}_{i,j}$.

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

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

$$Q(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}^{(l-1)}) = \text{const} - \frac{N}{2}\log\det(\boldsymbol{\Sigma}) -$$
$$- \sum_{n=1}^{N} \frac{1}{2}\text{tr}\Big((\widehat{\mathbf{x}}^{(n)(l-1)} - \mu)(\widehat{\mathbf{x}}^{(n)(l-1)} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}\Big) - \frac{1}{2}\text{tr}(\mathbf{C}\boldsymbol{\Sigma}^{-1}),$$

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

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

and

$$\widehat{\boldsymbol{\Sigma}}^{(l)} = \frac{1}{N}\sum_{n=1}^{N} \big[(\widehat{\mathbf{x}}^{(n)(l-1)} - \boldsymbol{\mu})(\widehat{\mathbf{x}}^{(n)(l-1)} - \boldsymbol{\mu})^\top + \mathbf{C}^{(n)}\big].$$

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!

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 harder to optimize directly
- no linearity in log-likelihood $\Rightarrow$ no imputation, more effort to compute expected likelihood

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 time
- Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data*. 3rd Edition.

Assignment 5 [5 %]

Simulate data (using Manual 9) from the mixture of two Gaussian distributions and implement the EM algorithm from Example 2 above. Use absolute change in ℓ of observed data as convergence criterion.

The following should naturally be done, though it is not mandatory:

- Visualize the resulting parametric density estimators.
- Try running your algorithm from different starting points.
  - How sensitive is your algorithm to your choice of starting point?
  - Can you find a bad starting point where your algorithm fails?

Project Ideas

- Cross-validation for PCA
  - A simulation study to compare the advantages of EM compared to what we did last week.
  - Two more approaches to CV for PCA to be considered and compared.
- Comparison of local regression implementations in different R packages.
  - Wickham (2011) examines several packages for KDE in R. Not only there are huge differences in terms of speed, but some of the packages are even inconsistent! Makes one wonder what is the situation with local regression.
- Various simulation studies on bandwidth selection in KDE or local linear regression.
- Comparison of variable selectors in regression.
  - Hastie et al. (2020) have some surprising results in their simulation study, but one important method (adaptive lasso) is omitted. Try to recreate the study with adaptive lasso included.