Computational Statistics II

Unit C.1: Missing data problems, Gibbs sampling and the EM algorithm

Tommaso Rigon

University of Milano-Bicocca

Ph.D. in Economics, Statistics and Data Science

Unit C.1

Main concepts

- Missing data problems;
- Data augmentation and Gibbs sampling;
- The EM algorithm and generalizations;
- Minorize maximize (MM) algorithms.

Main references

- Bishop, C. M. (2006). Pattern Recognition and Machine Learning, Chapter 9. Springer.
- Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *JRSS-B*, **39**(1), 1–38.
- Hunter, D. R., and Lange, K. (2004). A Tutorial on MM Algorithms. *The American Statistician*, **58**(1), 30–37.
- McLachlan, G. J. and Krishnan, T. (1998). The EM Algorithm and Extensions. Wiley.
- Robert, C. P., and Casella, G. (2009). Introducing Monte Carlo methods with R. Springer.

Missing data problems

- In this unit, we will take advantage of specific structures of the model to facilitate both frequentist and Bayesian computations via the EM and Gibbs sampling.
- In most cases, this will involve the introduction of **hidden features** of the model, sometimes called **latent variables**.
- Depending on the context, these latent quantities will have a precise meaning, or they will be regarded as purely abstract objects.
- An obvious example of latent components with a precise interpretation is the case of **missing** or **censored observations**.
- **Key idea**. If the complete data were available, computations would be easier. Besides, imputing the missing values could be interesting on its own.

Example: survival analysis with an exponential model

- Let $\boldsymbol{z} = (z_1, \dots, z_n)^\intercal$ be iid exponential random variables with rate parameter $\theta > 0$.
- If the prior $\theta \sim \text{Ga}(a, b)$, then thanks to conjugacy we get the following posterior

$$(\theta \mid \boldsymbol{z}) \sim \text{Ga}\left(a + n, b + \sum_{i=1}^n z_i\right).$$

- However, in many cases observations are **censored**, as in **Unit A.1**. In fact, we observe the values $\boldsymbol{t} = (t_1, \dots, t_n)^\intercal$ which are either complete $(t_i = z_i)$ or censored $(t_i \le z_i)$.
- If the observations were all **complete**, then inference would be straightforward.
- Intuitively, we aim at **sampling** or imputing the **missing information** from the appropriate conditional distribution to make inference about θ.

Data augmentation

- Let $\boldsymbol{X}$ be the **observed** data, following some distribution $\pi(\boldsymbol{X} \mid \boldsymbol{\theta})$, i.e. the **likelihood**, with $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ being an unknown set of parameters.
- Let $\pi(\boldsymbol{\theta})$ be the prior distribution associated to $\boldsymbol{\theta}$ and let $\pi(\boldsymbol{\theta} \mid \boldsymbol{X})$ be the posterior.
- Let $\boldsymbol{z} \in \mathcal{Z} \subseteq \mathbb{R}^q$ be a vector of **latent variables**, which are not observed.
- We assume that the likelihood function $\pi(\boldsymbol{X} \mid \boldsymbol{\theta})$ can be written as the marginal distribution of a **complete likelihood**, namely

$$\pi(\boldsymbol{X} \mid \boldsymbol{\theta}) = \int_{\mathcal{Z}} \pi(\boldsymbol{X}, \boldsymbol{z} \mid \boldsymbol{\theta}) \mathrm{d}\boldsymbol{z}.$$

- **Remark**. We focus on continuous densities w.r.t. the Lebesgue measure for notational simplicity, but these ideas apply in general.

Data augmentation

- The quantity $\pi(\boldsymbol{X}, \boldsymbol{z} \mid \boldsymbol{\theta})$ is the **complete** or **augmented** likelihood.

- Within the Bayesian framework, we treat the latent variables $\boldsymbol{z}$ as if they were an additional set of unknown parameters, leading to the **augmented posterior**

$$\pi(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{X}) \propto \pi(\boldsymbol{X}, \boldsymbol{z} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

- In other words, we aim at sampling $(\boldsymbol{\theta}^{(r)}, \boldsymbol{z}^{(r)})$ using MCMC from the joint posterior $\pi(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{X})$, which can be performed using any of the strategies we have described.

- If one is interested only in the original parameters $\boldsymbol{\theta}$ or in the latent dimensions $\boldsymbol{z}$, then it suffices to **ignore** the other set of parameters.

- We sample from $\pi(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{X})$ and then discard $\boldsymbol{z}$ rather than directly targeting $\pi(\boldsymbol{\theta} \mid \boldsymbol{X})$ because the **augmented likelihood** is typically **more tractable** than the original one.

Data augmentation schemes

- Unfortunately, there are **no general recipes** for finding useful data augmentation schemes. We will see proposals in the probit and logit case in **unit C.2**.

- In principle, whenever the likelihood can be expressed in an integral form, this leads to a potential data augmentation mechanism.

- The resulting augmented likelihood must be tractable, otherwise the whole procedure is of little practical utility.

- **Mixture models** greatly benefit from data-augmentation schemes, but we do not discuss them here because they would deserve an entire course on their own.

Data augmentation and Gibbs sampling

- Although in principle any MCMC strategy could be used to target $\pi(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{X})$, the Gibbs sampling is a natural choice in this setting.

- In fact, it is often the case that the following **full conditional distributions** are available in closed form. Moreover, they also have a nice interpretation.

- **Step 1**. Sample from the "posterior" of $\boldsymbol{\theta}$ based on the complete likelihood, namely

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{z}) \propto \pi(\boldsymbol{X}, \boldsymbol{z} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

- **Step 2**. Impute the missing observations $\boldsymbol{z}$ by sampling from the full conditional

$$\pi(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta}) \propto \pi(\boldsymbol{X}, \boldsymbol{z} \mid \boldsymbol{\theta}).$$

- Obviously, we are allowed to split $\boldsymbol{\theta}$ and $\boldsymbol{z}$ into blocks of parameters if this facilitates the Gibbs sampling.

Example: survival analysis with an exponential model

- Recall the exponential model example with censored data $\boldsymbol{t}$ and censorship indicators $\boldsymbol{d} = (d_1, \dots, d_n)^\intercal$. The **original likelihood** is therefore equal to

$$\pi(\boldsymbol{t}, \boldsymbol{d} \mid \theta) = \theta^{n_c} \exp\left\{-\theta \sum_{i=1}^n t_i\right\}, \qquad n_c = \sum_{i=1}^n d_i.$$

- **Remark**. This is a toy example. Indeed, under a Gamma prior, the posterior distribution of θ, based on the original likelihood, is known.

- In this setting, the latent variables $\boldsymbol{z}$ represent the complete survival times having exponential distribution so that the **complete likelihood** is

$$\pi(\boldsymbol{z} \mid \theta) = \theta^{n} \exp\left\{-\theta \sum_{i=1}^n z_i\right\}.$$

- The **Gibbs sampling** alternates between the Gamma full conditional $\pi(\boldsymbol{\theta} \mid \boldsymbol{z})$ and a sampling step from $\pi(\boldsymbol{z} \mid \boldsymbol{t}, \theta)$. Note that $(z_i - t_i \mid t_i, d_i, \theta) \overset{\text{ind}}{\sim} \text{Exp}(\theta)$ when $d_i = 0$.

The EM algorithm

- A Gibbs sampling based on data augmentation strategies is strongly connected with the so-called **expectation-maximization** (EM) algorithm.

- The EM is a deterministic algorithm that aims at **maximizing** the likelihood (MLE) or the posterior distribution (MAP), namely at finding

$$\arg\max_{\boldsymbol{\theta} \in \Theta} \pi(\boldsymbol{\theta} \mid \boldsymbol{X}) = \arg\max_{\boldsymbol{\theta} \in \Theta} \pi(\boldsymbol{X} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

- The EM is widely used within the frequentist and the Bayesian framework. The MLE case is recovered whenever $\pi(\boldsymbol{\theta}) \propto 1$.

- Compared to other gradient-based maximizers, it leads to a **monotonic sequence**. The target function always increases during the procedure, thus being more stable.

- On the other hand, the EM **requires** a (tractable) **augmented likelihood**. Moreover, the EM could be slower than other algorithms to reach convergence.

The EM algorithm

- The EM algorithm alternates between the following steps, which are reminiscent of those of the Gibbs sampling, as they involve similar quantities.

- Initialize the algorithm at a reasonable $\boldsymbol{\theta}^{(0)}$. The generic iteration proceeds as follows.

- **Step 1 (Expectation)**. Let $\boldsymbol{\theta}^{(r)}$ be the current value of the maximization procedure, then obtain the function

$$\mathcal{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(r)}) = \mathbb{E}\{\log \pi(\boldsymbol{X}, \boldsymbol{z} \mid \boldsymbol{\theta})\},$$

where the expectation is taken with respect to the conditional law $\pi(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{(r)})$.

- **Step 2 (Maximization)**. The new value of the procedure $\boldsymbol{\theta}^{(r+1)}$ is obtained by maximizing the function

$$\arg\max_{\boldsymbol{\theta} \in \Theta} \mathcal{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(r)}) + \log \pi(\boldsymbol{\theta}).$$

- In many cases, the E-step amounts at calculating $\mathbb{E}(\boldsymbol{z})$ and then plugging-in this quantity in the augmented log-likelihood. Indeed, $\log \pi(\boldsymbol{X}, \boldsymbol{z} \mid \boldsymbol{\theta})$ is often linear in $\boldsymbol{z}$.

Example: survival analysis with an exponential model

- Recall that in the exponential model example, we have that $(z_i - t_i \mid t_i, d_i, \theta) \overset{\text{ind}}{\sim} \text{Exp}(\theta)$ when $d_i = 0$ and the augmented likelihood is $\pi(\boldsymbol{z} \mid \theta) = \theta^n \exp\{-\theta \sum_{i=1}^n z_i\}$.

- Let us focus on the maximum likelihood so that $\pi(\theta) \propto 1$.

- **Step 1 (Expectation)**. Let $\theta^{(r)}$ be the current value of the procedure, then

$$\mathcal{Q}(\theta \mid \theta^{(r)}) = n \log \theta - \theta \sum_{i=1}^n \mathbb{E}(z_i) = n \log \theta - \theta \sum_{i=1}^n \left\{ t_i + \frac{(1 - d_i)}{\theta^{(r)}} \right\},$$

  where the expectation is taken with respect to the conditional law $\pi(\boldsymbol{z} \mid \boldsymbol{t}, \boldsymbol{d}, \theta^{(r)})$.

- **Step 2 (Maximization)**. The new value of the procedure $\boldsymbol{\theta}^{(r+1)}$ is obtained by considering the maximum of $\mathcal{Q}(\theta \mid \theta^{(r)})$, thus obtaining

$$\theta^{(r+1)} = \left(\frac{1}{n} \sum_{i=1}^n t_i + \frac{n - n_c}{n} \frac{1}{\theta^{(r)}} \right)^{-1}.$$

Why does the EM work?

Theorem (monotonic EM sequence)

The EM sequence for finding the MLE satisfies the following inequality

$$\pi(\boldsymbol{X} \mid \boldsymbol{\theta}^{(r+1)}) \geq \pi(\boldsymbol{X} \mid \boldsymbol{\theta}^{(r)}).$$

Similarly, the EM sequence for finding the MAP satisfies the following inequality

$$\pi(\boldsymbol{\theta}^{(r+1)} \mid \boldsymbol{X}) \geq \pi(\boldsymbol{\theta}^{(r)} \mid \boldsymbol{X}).$$

- With some further continuity assumptions w.r.t. $\boldsymbol{\theta}$, this theorem implies that the EM is guaranteed to reach a **stationary point**.
- If the posterior / likelihood function is concave, the stationary point will also be the global maximum.
- In general, as in any maximization procedure, it is recommended to initialize the algorithm at different starting points.

Sketch of the proof

- In the first place, recognize that the following identity holds (do it as an exercise!)

$$\log \pi(\boldsymbol{\theta} \mid \boldsymbol{X}) = \log \pi(\boldsymbol{X}, \boldsymbol{z} \mid \boldsymbol{\theta}) + \log \pi(\boldsymbol{\theta}) - \log \pi(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta}) - \log \pi(\boldsymbol{X}),$$

  Consequently, one gets the following identity

$$\log \pi(\boldsymbol{\theta} \mid \boldsymbol{X}) = \mathcal{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}') + \log \pi(\boldsymbol{\theta}) - \mathbb{E}\{\log \pi(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta})\} - \log \pi(\boldsymbol{X}),$$

  after taking the expectation w.r.t. $\pi(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta}')$.

- Let $\boldsymbol{\theta}^{(r)}$ and $\boldsymbol{\theta}^{(r+1)}$ be subsequent steps in the EM procedure. Then, it necessarily holds that

$$\mathcal{Q}(\boldsymbol{\theta}^{(r+1)} \mid \boldsymbol{\theta}^{(r)}) + \log \pi(\boldsymbol{\theta}^{(r+1)}) \geq \mathcal{Q}(\boldsymbol{\theta}^{(r)} \mid \boldsymbol{\theta}^{(r)}) + \log \pi(\boldsymbol{\theta}^{(r)}),$$

  as the value $\boldsymbol{\theta}^{(r+1)}$ is indeed maximizing the left-hand-side. Furthermore, note that because of Jensen's inequality, we get

$$\mathbb{E}\left\{\log \frac{\pi(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{(r+1)})}{\pi(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{(r)})}\right\} \leq \log \mathbb{E}\left\{\frac{\pi(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{(r+1)})}{\pi(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{(r)})}\right\} = 0,$$

  expectations being taken w.r.t. to $\pi(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{(r)})$. This implies that

$$-\mathbb{E}\{\log \pi(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{(r+1)})\} \geq -\mathbb{E}\{\log \pi(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{(r)})\}.$$

- The proof follows by combining the above results, after noting that

$$\begin{aligned}\log \pi(\boldsymbol{\theta}^{(r+1)} \mid \boldsymbol{X}) &= \mathcal{Q}(\boldsymbol{\theta}^{(r+1)} \mid \boldsymbol{\theta}^{(r)}) + \log \pi(\boldsymbol{\theta}^{(r+1)}) - \mathbb{E}\{\log \pi(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{(r+1)})\} - \log \pi(\boldsymbol{X}) \geq \\ &\geq \mathcal{Q}(\boldsymbol{\theta}^{(r)} \mid \boldsymbol{\theta}^{(r)}) + \log \pi(\boldsymbol{\theta}^{(r)}) - \mathbb{E}\{\log \pi(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{(r)})\} - \log \pi(\boldsymbol{X}) = \log \pi(\boldsymbol{\theta}^{(r)} \mid \boldsymbol{X}).\end{aligned}$$

An alternative derivation of the EM

- There exists an alternative derivation of the EM purely based on **maximization**.
- Albeit less common, this way of thinking leads to a more **elegant proof** and puts the basis for variational Bayes (VB) procedures **unit D.1**.
- Let $q(\boldsymbol{z}) \in \mathbb{Q}$ be a generic density of the latent variables $\boldsymbol{z}$ and define

$$\mathcal{L}\{q(\boldsymbol{z}) \mid \boldsymbol{X}, \boldsymbol{\theta}\} = \mathbb{E}_q\left(\log \frac{\pi(\boldsymbol{X}, \boldsymbol{z} \mid \boldsymbol{\theta})}{q(\boldsymbol{z})}\right),$$

  where the expectations are taken w.r.t. $q(\boldsymbol{z})$.
- Moreover, define the **Kullback-Leibler divergence**

$$\text{KL}\{q(\boldsymbol{z}) \mid\mid \pi(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta})\} = -\mathbb{E}_q\left(\log \frac{\pi(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta})}{q(\boldsymbol{z})}\right).$$

A maximization / maximization procedure

- Let us focus on the MLE case for notational simplicity. The MAP case is recovered with minor adjustments (do it as an exercise!)

- For any $q \in \mathbb{Q}$ the following **identity** holds true

$$\log \pi(\boldsymbol{X} \mid \boldsymbol{\theta}) = \mathcal{L}\{q(\boldsymbol{z}) \mid \boldsymbol{X}, \boldsymbol{\theta}\} + \text{KL}\{q(\boldsymbol{z}) \mid\mid \pi(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta})\}.$$

- Since the Kullback-Leibler divergence $\text{KL}\{q(\boldsymbol{z}) \mid\mid \pi(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta})\} \geq 0$, then we will have

$$\mathcal{L}\{q(\boldsymbol{z}) \mid \boldsymbol{X}, \boldsymbol{\theta}\} \leq \log \pi(\boldsymbol{X} \mid \boldsymbol{\theta}),$$

meaning that $\mathcal{L}\{q(\boldsymbol{z}) \mid \boldsymbol{\theta}, \boldsymbol{X}\}$ is the **lower bound** of the log-likelihood.

- This suggests that the MLE can be found **maximizing the lower bound**, since

$$\arg\max_{\boldsymbol{\theta} \in \Theta} \log \pi(\boldsymbol{X} \mid \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta} \in \Theta} \max_{q \in \mathbb{Q}} \mathcal{L}\{q(\boldsymbol{z}) \mid \boldsymbol{X}, \boldsymbol{\theta}\}.$$

- Indeed, the value $q(\boldsymbol{z}) = \pi(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta})$ is the maximum of $\mathcal{L}\{q(\boldsymbol{z}) \mid \boldsymbol{X}, \boldsymbol{\theta}\}$, because

$$\mathcal{L}\{q(\boldsymbol{z}) \mid \boldsymbol{X}, \boldsymbol{\theta}\} = \log \pi(\boldsymbol{X} \mid \boldsymbol{\theta}) - \underbrace{\text{KL}\{q(\boldsymbol{z}) \mid\mid \pi(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta})\}}_{=0} = \log \pi(\boldsymbol{X} \mid \boldsymbol{\theta}).$$

A maximization / maximization procedure

- Consequently, the MLE can be obtained by iteratively maximizing $\mathcal{L}\{q(\boldsymbol{z}) \mid \boldsymbol{\theta}, \boldsymbol{X}\}$ over $q(\boldsymbol{z})$ for a given value of $\boldsymbol{\theta}$ and then over $\boldsymbol{\theta}$ for a given $q(\boldsymbol{z})$.

- Let $\boldsymbol{\theta}^{(r)}$ be the current value of the procedure.

- **Step 1 (Maximization over q)**. Given the fixed value $\boldsymbol{\theta}^{(r)}$, obtain

$$\pi(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{(r)}) = \arg\max_{q \in \mathbb{Q}} \mathcal{L}\{q(\boldsymbol{z}) \mid \boldsymbol{X}, \boldsymbol{\theta}^{(r)}\} = \arg\min_{q \in \mathbb{Q}} \mathrm{KL}\{q(\boldsymbol{z}) \mid\mid \pi(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{(r)})\}.$$

- **Step 2 (Maximization over θ)**. Given the locally optimal value $q(\boldsymbol{z}) = \pi(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{(r)})$, obtain the new value $\boldsymbol{\theta}^{(r+1)}$ as the maximizer

$$\boldsymbol{\theta}^{(r+1)} = \arg\max_{\boldsymbol{\theta} \in \Theta} \mathcal{L}\{\pi(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{(r)}) \mid \boldsymbol{X}, \boldsymbol{\theta}\} = \arg\max_{\boldsymbol{\theta} \in \Theta} \mathcal{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(r)}).$$

- These are the steps of the EM, which therefore has an alternative interpretation.

- Moreover, recalling that $\mathcal{L}\{\pi(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{(r)}) \mid \boldsymbol{X}, \boldsymbol{\theta}^{(r)}\} = \log \pi(\boldsymbol{X} \mid \boldsymbol{\theta}^{(r)})$, the monotonicity property of the EM is obvious.

Generalizations of the EM

- Sometimes the **maximization** of $\mathcal{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(r)}) + \log \pi(\boldsymbol{\theta})$, namely the maximization step, could be **difficult**.

- Thus, an obvious generalization of the EM algorithm that preserves the monotonicity of the procedure is considering some value $\boldsymbol{\theta}^{(r+1)}$ such that

$$\mathcal{Q}(\boldsymbol{\theta}^{(r+1)} \mid \boldsymbol{\theta}^{(r)}) + \log \pi(\boldsymbol{\theta}^{(r+1)}) \geq \mathcal{Q}(\boldsymbol{\theta}^{(r)} \mid \boldsymbol{\theta}^{(r)}) + \log \pi(\boldsymbol{\theta}^{(r)})$$

that is, $\boldsymbol{\theta}^{(r+1)}$ **increases the function** rather maximizing it.

- An example is the **expectation conditional maximization** (ECM) of Meng and Rubin (1993), where the parameters are partitioned into sub-groups and iteratively maximized.

- Similar ideas can be applied to generalize the expectation step by doing a "partial" update in the maximization of q.

MM algorithms

- We finally consider a large class of optimization methods called **minorize maximize** (MM) that includes the EM as a special case.

- MM methods do not involve missing data or data augmentations but rather rely on general **convexity** arguments.

- The MM is used to optimize a $\ell(\boldsymbol{\theta}; \boldsymbol{X})$ of the parameters $\boldsymbol{\theta}$ and the data $\boldsymbol{X}$, with $f(\cdot)$ being the posterior distribution, the likelihood, or a general loss function.

- Let $\boldsymbol{\theta}^{(r)}$ be the current value of the iterative maximization procedure. We are seeking for a **minorization function** $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(r)})$, such that

$$g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(r)}) \leq \ell(\boldsymbol{\theta}; \boldsymbol{X}), \qquad \text{for any } \boldsymbol{\theta} \in \Theta,$$

and satisfying $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}) = \ell(\boldsymbol{\theta}; \boldsymbol{X})$.

MM algorithms

- In MM algorithms we iteratively maximize the **lower bound** $g(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}, \boldsymbol{X})$, so that

$$\boldsymbol{\theta}^{(r+1)} = \arg\max_{\boldsymbol{\theta} \in \Theta} g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(r)})$$

- MM leads to **monotonic sequences**, since

$$\ell(\boldsymbol{\theta}^{(r+1)}; \boldsymbol{X}) \geq g(\boldsymbol{\theta}^{(r+1)} \mid \boldsymbol{\theta}^{(r)}) \geq g(\boldsymbol{\theta}^{(r)} \mid \boldsymbol{\theta}^{(r)}) = \ell(\boldsymbol{\theta}^{(r)}; \boldsymbol{X}).$$

- This property ensures remarkable numerical stability but does not provide any hint about the actual construction of $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(r)})$.

- The EM is indeed a **special case** of this framework, recovered in the MLE case by defining

$$g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(r)}) = \mathcal{L}\{\pi(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{(r)}) \mid \boldsymbol{X}, \boldsymbol{\theta}\} \leq \log \pi(\boldsymbol{X} \mid \theta).$$

  and recalling that $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(r)}) = \mathcal{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(r)}) + \text{const}$, and that $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}) = \log \pi(\boldsymbol{X} \mid \boldsymbol{\theta})$.

- We will see an example in **unit C.2** for the logistic regression case.