

Computational Statistics II

Unit C.1: Missing data problems

Tommaso Rigon

University of Milano-Bicocca

Ph.D. in Economics and Statistics



Main concepts

- Missing data problems;
- Data augmentation and Gibbs sampling;
- Connection with the EM algorithm.

Missing data problems

- In this unit we will take advantage of specific structures of the model to facilitate Bayesian computations.
- However, in most cases, this will involve the introduction of **hidden features** of the model, sometimes called **latent variables**, which might be interesting per se.
- Depending on the context, these latent quantities will have a precise meaning or they will be regarded as purely abstract objects.
- An obvious examples 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 $\mathbf{z} = (z_1, \dots, z_n)^\top$ be iid exponential random variables with rate parameter θ .
- If the prior $\theta \sim \text{Ga}(a, b)$, then thanks to conjugacy we get the following posterior

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

- However, in many cases observations are **censored**, as in **Unit A.1**. In fact, we observe the values $\mathbf{t} = (t_1, \dots, t_n)^\top$ which are either complete ($t_i = z_i$) or censored ($t_i \leq z_i$).
- If the observations were all **complete**, then inference would be straightforward.
- Intuitively, what we are going to do is to **sample** the **missing information** from the corresponding conditional distribution in order to make inference about θ .

Data augmentation

- Let \mathbf{X} be the **observed** data, following some distribution $\pi(\mathbf{X} \mid \theta)$, i.e. the **likelihood**, with $\theta \in \Theta \subseteq \mathbb{R}^p$ being an unknown set of parameters.
- Let $\pi(\theta)$ be the prior distribution associated to θ and let $\pi(\theta \mid \mathbf{X})$ be the posterior.
- Let $\mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^q$ be a vector of **latent variables**.
- We assume that the likelihood function $\pi(\mathbf{X} \mid \theta)$ can be written as the marginal distribution of a **complete likelihood**, namely

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

- **Remark.** We focus on continuous density w.r.t. the Lebesgue measure for the sake of notational simplicity, but the same idea applies more generally.

Data augmentation

- The quantity $\pi(\mathbf{X}, \mathbf{z} \mid \boldsymbol{\theta})$ is the **complete** or **augmented** likelihood.
- Within the Bayesian framework, we should treat the latent variables \mathbf{z} as if they were an additional set of unknown parameters, leading to the **augmented posterior**

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

- In other words, we aim at sampling $(\boldsymbol{\theta}^{(r)}, \mathbf{z}^{(r)})$ using MCMC from the joint posterior $\pi(\boldsymbol{\theta}, \mathbf{z} \mid \mathbf{X})$, 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 \mathbf{z} , then it suffices to **ignore** the other set of parameters.
- If the interest is in $\boldsymbol{\theta}$, the advantage of sampling from $\pi(\boldsymbol{\theta}, \mathbf{z} \mid \mathbf{X})$ and then discarding \mathbf{z} rather than directly targeting $\pi(\boldsymbol{\theta} \mid \mathbf{X})$ is that the **augmented likelihood** is typically **more tractable** than the original one.

Data augmentation and Gibbs sampling

- Although in principle any MCMC strategy could be used to target $\pi(\theta, \mathbf{z} \mid \mathbf{X})$, the Gibbs sampling is often a natural choice.
- 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 θ based on the complete likelihood, namely

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

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

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

- Obviously, we are allowed to split θ and \mathbf{z} into blocks of parameters if this facilitate the Gibbs sampling.

Example: survival analysis with an exponential model

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

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

- **Remark.** This is a toy example whose purpose is fixing ideas. Indeed, under a Gamma prior, the posterior distribution of θ using this likelihood is also available.
- In this setting, the latent variables \mathbf{z} represent the complete survival times having exponential distribution, so that the **complete likelihood** is

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

- The Gibbs sampling therefore alternates the full conditional $\pi(\theta \mid \mathbf{z})$ and an imputation sampling step from $\pi(\mathbf{z} \mid \mathbf{t}, \theta)$. As for the latter, note that $(z_i - t_i \mid t_i, \theta) \stackrel{\text{ind}}{\sim} \text{Exp}(\theta)$.

Connection with the EM algorithm (recap)

- 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).
- Hence, as opposed to sampling strategies, the EM is widely used both within the frequentist and the Bayesian framework.
- Compared to other gradient-based maximizers, it leads to a monotonic sequence \implies the function always increases during the procedure.
- On the other hand, it requires a (tractable) augmented likelihood.

The EM algorithm (recap)

- The EM algorithm alternates between the following steps, which are reminiscent of those of the Gibbs sampling, as they involve similar quantities.
- **Step 1 (Expectation)**. Let $\theta^{(r)}$ be the current value of the maximization procedure, then obtain the function

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

where the expectation is taken with respect to the conditional distribution $\pi(\mathbf{z} \mid \mathbf{X}, \theta)$.

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

$$\arg \max_{\theta \in \Theta} Q(\theta \mid \theta^{(r)}).$$

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

Data augmentation schemes

- Differently from other MCMC methods, there is **no general recipe** for finding useful data augmentation schemes.
- In principle, whenever the likelihood can be expressed in an integral form, this leads to a potential data augmentation mechanism.
- However, the resulting augmented likelihood has to be tractable (in some sense), otherwise the whole procedure is of little practical utility.
- **Mixture models** greatly benefit from data-augmentation schemes, but we do not discuss them because they would deserve an entire course on their own.