Computational Statistics II

Unit C.1: Missing data problems

Tommaso Rigon

University of Milano-Bicocca

Ph.D. in Economics and Statistics

Unit C.1

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 $\boldsymbol{z} = (z_1, \dots, z_n)^\intercal$ 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

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

- 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, 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 $\boldsymbol{X}$ 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**.

- 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 density w.r.t. the Lebesgue measure for the sake of notational simplicity, but the same idea applies more generally.

Data augmentation

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

- Within the Bayesian framework, we should 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.

- If the interest is in $\boldsymbol{\theta}$, the advantage of sampling from $\pi(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{X})$ and then discarding $\boldsymbol{z}$ rather than directly targeting $\pi(\boldsymbol{\theta} \mid \boldsymbol{X})$ 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 $\pi(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{X})$, 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 $\boldsymbol{\theta}$ 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 facilitate 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 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 $\boldsymbol{z}$ 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 therefore alternates the full conditional $\pi(\boldsymbol{\theta} \mid \boldsymbol{z})$ and an imputation sampling step from $\pi(\boldsymbol{z} \mid \boldsymbol{t}, \theta)$. As for the latter, note that $(z_i - t_i \mid t_i, \theta) \overset{\text{ind}}{\sim} \text{Exp}(\theta)$.

Connection with 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).

- 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 $\boldsymbol{\theta}^{(r)}$ be the current value of the maximization procedure, then obtain the function

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

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

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

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

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

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 pointless.
- A crucial class of models that greatly benefit from data-augmentation schemes are **mixture models**, which are because they indeed deserve an entire course on their own.