Bayesian statistics

EM algorithm and mixture models

Leonardo Egidi

2024/2025

Università di Trieste

Table of contents i

Missing-data models

Missing-data models i

- Missing-data models cover many statistical settings, including censoring models and mixture and latent variable models (tobit, probit, stochastic volatility, etc.).
- The representation is as follows:

$$p(y|\theta) = \int f(y, z|\theta) dz, \tag{1}$$

  where z refers to the *missing/augmented* data, and $(y, z) \sim f(y, z|\theta)$.
- In many applications, the vector z merely serves to simplify calculations, as it does not necessarily have a specific meaning for the corresponding statistical problem.
- The model (1) can be seen as a missing-data model in the sense that z can be interpreted as missing from the observations y.

Missing-data models ii

- We refer to the function $L^c(\theta; y, z) = f(y, z|\theta)$ as the *complete model* or *complete-data likelihood*, which is the likelihood we would obtain were we to observe (y, z), also called the complete data.
- In a mixture model with J components the likelihood has the form:

$$p(y|\boldsymbol{\theta}, \phi) = \prod_{i=1}^{n} \sum_{j=1}^{J} \phi_j p(y_i|\theta_j), \tag{2}$$

where ϕ_j is the j-th group-membership assignment probability, and $p(y_i|\theta_j)$ is the likelihood of the i-th data point in group j.

Missing-data models iii

- By introducing a vector of auxiliary/latent indicators, $z = (z_1, z_2, \ldots, z_n)$, such that $\phi_j = \Pr(z_i = j)$, and given that $y_i|z_i = j \sim p(y_i|\theta_j)$, we can define a complete likelihood for (y_i, z_i) as follows:

$$L^c(\theta, \phi; y_i, z_i) \propto f(y_i, z_i|\theta, \phi) = \phi_j p(y_i|\theta_j), \tag{3}$$

  and summing over the J group we can resemble Equation (1) and re-find Equation (2) according to a missing-data representation.

The expectation-maximization (EM) algorithm

The EM algorithm i

- The EM algorithm is a deterministic optimization technique that takes advantages of the representation (1) to find the parameters' values which maximize an expected version of the complete-data likelihood.
- Assume to observe $y = (y_1, \ldots, y_n)$ jointly distributed from $p(y|\theta)$ that satisfies the representation in (1). We want to compute $\hat{\theta} = \text{argmax } L(\theta; y) = \text{argmax } p(y|\theta)$.
- We can derive a useful relationship by using the conditional laws:

$$f(y, z|\theta) = f(z|y, \theta)p(y|\theta), \tag{4}$$

then we apply the logarithmic transformation, moving $\log p(y|\theta)$ to the left side:

$$\begin{aligned} \log p(y|\theta) &= \log f(y, z|\theta) - \log f(z|y, \theta) \\ &= \log L^c(\theta; y, z) - \log f(z|y, \theta). \end{aligned} \tag{5}$$

The EM algorithm ii

- Suppose now to take expectations of both sides, treating z as a random variable. Thus, for any value θ_0, we take the expectation with respect to the distribution $f(z|\theta_0, y)$:

$$\log p(y|\theta) = \mathrm{E}_{\theta_0}[\log L^c(\theta; y, z)] - \mathrm{E}_{\theta_0}[\log f(z|y, \theta)], \tag{6}$$

  where the first term does not depend on z.

- In the EM algorithm, while we aim at maximizing $\log p(y|\theta)$, only the first term on the right side of (6) will be considered. We denote

$$Q(\theta|\theta_0, y) = \mathrm{E}_{\theta_0}[\log L^c(\theta; y, z)]. \tag{7}$$

The EM algorithm iii

- The EM algorithm indeed proceeds iteratively by maximizing $Q(\theta|\theta_0, y)$ at each iteration, and, if $\hat{\theta}_{(1)}$ is the value of θ maximizing $Q(\theta|\theta_0, y)$, by replacing θ_0 by the updated value $\hat{\theta}_{(1)}$. In this manner, a sequence of estimators $\{\hat{\theta}_{(j)}\}_j$ is obtained, where $\hat{\theta}_{(j)}$ is the value of θ maximizing $Q(\theta|\hat{\theta}_{(j-1)}, y)$.
- The iterative scheme thus contains both an expectation step and a maximization step, giving the algorithm its name. See details in Algorithm 1.

The EM algorithm iv

Algorithm 1: The EM algorithm

Input: complete likelihood $L^c(\theta; y, z)$, threshold ϵ
Output: sequence of estimators $\{\hat{\theta}_{(j)}\}_j$
Initialize: set $j = 0$, pick a value θ_0
while $||\hat{\theta}_{(j)} - \hat{\theta}_{(j-1)}|| > \epsilon$ **do**

1. Compute (*the E-step*))

$$Q(\theta|\hat{\theta}_{(j)}, y) = \mathrm{E}_{\hat{\theta}_{(j)}}[\log L^c(\theta; y, z)],$$

where the expectation is with respect to $f(z|\hat{\theta}_{(j)}, y)$

2. Maximize $Q(\theta|\hat{\theta}_{(j)}, y)$ in θ and take (*the M-step*)

$$\hat{\theta}_{(j+1)} = \underset{\theta}{\mathrm{argmax}} Q(\theta|\hat{\theta}_{(j)}, y)$$

and set $j = j + 1$
end
return $\hat{\theta}_{(j+1)}$

The EM algorithm v

- By virtue of Jensen's inequality, it is easy to show that , at each step of the EM algorithm, the likelihood on the left side of Equation (6) increases,

$$L(\hat{\theta}_{(j+1)}; y) \geq L(\hat{\theta}_{(j)}; y).$$

  This means that under some conditions every limit point of an EM sequence $\{\hat{\theta}_{(j)}\}_j$ is a stationary point of $L(\theta; y)$, albeit not necessarily the maximum likelihood estimator or even a local maximum.

- It thus means that, in practice, running the EM algorithm several times with different, randomly chosen starting points is recommended if one wants to avoid using a poor approximation to the true maximum.

- Implementing the EM algorithm thus means being able to (a) compute the function $Q(\theta'|\theta, y)$ and (b) maximize this function.

EM for a Gaussian mixture i

- Suppose a two-component Gaussian mixture:

$$p(y|\mu) = \frac{1}{4}\mathcal{N}(\mu_1, 1) + \frac{3}{4}\mathcal{N}(\mu_2, 1), \tag{8}$$

where this likelihood is bimodal. This model can easily be expressed as a missing-data model, assume then that $(z_1, z_2, \ldots, z_n) \in \{1, 2\}^n$ such that:

$$\Pr(z_i = 1) = 1 - \Pr(z_i = 2) = 1/4, \quad y_i|z_i = j \sim \mathcal{N}(\mu_j, 1). \tag{9}$$

- The observed likelihood is then equal to

$$L(\mu; y) \propto \prod_{i:z_i=1} \frac{1}{4} \exp\{-(y_i - \mu_1)^2/2\} \prod_{i:z_i=2} \frac{3}{4} \exp\{-(y_i - \mu_2)^2/2\}. \tag{10}$$

EM for a Gaussian mixture ii

- We can then compute the expected complete-data log-likelihood (E-step) as:

$$Q(\theta'|\theta, y) = -\frac{1}{2}\sum_{i=1}^{n} \mathrm{E}_\theta[z_i(y_i - \mu_1)^2 + (1 - z_i)(y_i - \mu_2)^2|y] \tag{11}$$

- Solving the M-step provides the closed form expressions:

$$\begin{aligned} \mu_1' &= \mathrm{E}_\theta\left[\sum_{i=1}^{n} z_i y_i \Big| y\right] / \mathrm{E}_\theta\left[\sum_{i=1}^{n} z_i \Big| y\right] \\ \mu_2' &= \mathrm{E}_\theta\left[\sum_{i=1}^{n} (1 - z_i) y_i \Big| y\right] / \mathrm{E}_\theta\left[\sum_{i=1}^{n} (1 - z_i) \Big| y\right]. \end{aligned} \tag{12}$$

- See Example 5.15 in *Introducing Monte Carlo Methods with R* for a graphical inspection of the algorithm.

EM for Bayesian inference

EM for Bayesian inference i

- In problems with many parameters, Gaussian approximations to the joint distribution are often useless, and the joint mode is typically not helpful.
- It is often useful, however, to base an approximation on a marginal posterior model of a subset of the parameters: we use the notation $\theta = (z, \phi)$ and suppose we are interested in first approximating the marginal posterior $\pi(\phi|y)$.
- The EM algorithm just introduced can be viewed as an iterative method for finding the mode of the marginal posterior density, $\pi(\phi|y)$, and is extremely useful for many common models for which it is hard to maximize $\pi(\phi|y)$ directly but easy to work with $\pi(z|\phi, y)$ and $\pi(\phi|z, y)$.
- In what follows, we think of ϕ as the parameters in our problem and z as missing data: as for the frequentist/classical approach, EM is widely applicable when the models can be re-expressed as distributions on augmented parameter spaces z (such as mixture, hierarchical models, etc.)

EM for Bayesian inference ii

- Bayesian inference draws *no distinction between missing data and parameters*: both are uncertain, and they have a joint posterior distribution, conditional on observed data.
- As a final output, the EM finds the modes of $\pi(\phi|y)$ averaging over z.
- We start by resembling Equation (5) in the following way:

$$\log \pi(\phi|y) = \log \pi(z, \phi|y) - \log \pi(z|\phi, y), \tag{13}$$

and take expectations of both sides, treating z as a random variable with the distribution $\pi(z|\phi_0, y)$, where ϕ_0 is the current parameter guess.

- Thus, averaging over z yields:

$$\log \pi(\phi|y) = \mathrm{E}_{\phi_0}[\log \pi(z, \phi|y)] - \mathrm{E}_{\phi_0}[\log \pi(z|\phi, y)], \tag{14}$$

where the expectation is taken over z under the distribution $\pi(z|\phi_0, y)$, and the second term in (14) is maximized at $\phi = \phi_0$.

EM for Bayesian inference iii

- Analogously as the classical EM, we consider the first term of Equation (14) to maximize $\log \pi(\phi|y)$ and we denote it by

$$Q(\phi|\phi_0, y) = \mathrm{E}_{\phi_0}[\log \pi(z, \phi|y)], \tag{15}$$

  the expected log-posterior density function.

- Because the marginal posterior density, $\pi(\phi|y)$, increases in each step of the EM algorithm, and because the Q function is maximized at each step, EM converges to a local mode of the posterior density except in some special cases.

- An analogous version of the Algorithm 1 can be derived for the Bayesian case, just by using the new Q in (15) for the E and the M steps. Full details are provided in Algorithm 2.

- A simple way to search for multiple modes (as in mixtures) with EM is to start the iterations at many points throughout the parameter space.

EM for Bayesian inference iv

Algorithm 2: The Bayesian EM algorithm

Input: log-posterior density $\log \pi(z, \phi|y)$, threshold ϵ
Output: sequence of modes $\{\hat{\phi}_{(j)}\}_j$
Initialize: set $j = 0$, pick a value ϕ_0
while $||\hat{\phi}_{(j)} - \hat{\phi}_{(j-1)}|| > \epsilon$ **do**

1. Compute (*the E-step*))

$$Q(\phi|\hat{\phi}_{(j)}, y) = \mathrm{E}_{\hat{\phi}_{(j)}}[\log \pi(z, \phi|y)],$$

where the expectation is with respect to $\pi(z|\hat{\phi}_{(j)}, y)$

2. Maximize $Q(\theta|\hat{\phi}_{(j)}, y)$ in ϕ and take (*the M-step*)

$$\hat{\phi}_{(j+1)} = \underset{\phi}{\mathrm{argmax}} Q(\phi|\hat{\phi}_{(j)}, y)$$

and set $j = j + 1$
end
return $\hat{\phi}_{(j+1)}$

Example: Gaussian model with unknown mean and variance i

- Suppose we weigh an object on a scale n times, with weightings $(y_1, y_2, \ldots, y_n)$ assumed to be independent with a $\mathcal{N}(\mu, \sigma^2)$, where μ is the true weight of the object.
- Assume a prior $\mathcal{N}(\mu_0, \tau_0^2)$ on μ, with μ_0 and τ_0^2 known, and the standard noninformative uniform prior on $\log \sigma$.
- Because the model is not fully conjugate, we can use the EM algorithm to find the marginal posterior mode of μ, averaging over σ; that is, (μ, σ) corresponds to (ϕ, z) in the general notation.
- The joint log-posterior density is

$$\log p(\mu, \sigma|y) = -\frac{1}{2\tau_0^2}(\mu - \mu_0)^2 - (n+1)\log \sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 + \text{const}, \tag{16}$$

ignoring terms that do not depend on μ or σ^2.

Example: Gaussian model with unknown mean and variance ii

- *E-step*:

$$\begin{aligned} \mathrm{E}_{\mu'}[\log p(\mu, \sigma|y)] = & -\frac{1}{2\tau_0^2}(\mu - \mu_0)^2 - (n+1)\log \mathrm{E}_{\mu'}[\log \sigma] \\ & -\frac{1}{2}\mathrm{E}_{\mu'}\left[\frac{1}{\sigma^2}\right]\sum_{i=1}^{n}(y_i - \mu)^2 + \text{const.} \end{aligned} \tag{17}$$

We note that $\sigma^2|\mu, y \sim \text{Inv}\chi^2\left(n, \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2\right)$. Then the conditional posterior distribution of $1/\sigma^2$ is a scaled χ^2 with mean equal to $\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \mu')^2\right)^{-1}$, which can be supplied in (17). Notice that $\mathrm{E}_{\mu'}[\log \sigma]$ does not depend on μ and then is constant, will not affect the M-step.

Example: Gaussian model with unknown mean and variance iii

- *M-step*: we need to find out μ that maximizes

$$\begin{aligned} \mathrm{E}_{\mu'}[\log p(\mu, \sigma|y)] = & -\frac{1}{2\tau_0^2}(\mu-\mu_0)^2 - \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu')^2\right)^{-1} \times \\ & \sum_{i=1}^{n}(y_i-\mu)^2 + \text{const}, \end{aligned} \tag{18}$$

  that has the form of a normal log posterior density. Thus the M-step is achieved by the mode of the equivalent posterior density, which is:

$$\mu^* = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu')^2}\bar{y}}{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu')^2}}. \tag{19}$$

- If we iterate this computation, μ converges to the marginal mode of $\pi(\mu|y)$.

Sum-up

- The EM algorithm is a two-step optimization algorithm which is very beneficial when we have an augmented data structure and we may want to average over some missing data/auxiliary parameters z, both under a classical or a Bayesian approach.
- The EM algorithm has some similarities with the Gibbs sampling algorithm, which also enjoys the data augmentation representation, and uses the full conditionals to sample new values and approximate the underlying target distribution.
- Moreover, there exist some EM extensions, such as the Monte Carlo EM algorithm and the ECM algorithm (see BDA, chapter 13 for more details).
- However, the main connection of the EM algorithm is with Variational Inference (VI) methods, where the iterations lead to a closed-form approximation that is the closest fit to the posterior distribution within some specified class of functions.

Further reading

To properly capture EM (with examples):

- Sections 5.4.2 and 5.4.3 from *Introducing Monte Carlo methods with R*, by C. Robert and G. Casella.
- Section 13.4 from *Bayesian Data Analysis*, by A. Gelman et al.