Incomplete Data Analysis

Vanda Inácio

University of Edinburgh

Semester 1, 2020/2021

EM algorithm in exponential families

- The computation of the E and M steps simplify when it can be shown that the log likelihood of the complete data is linear in the **sufficient statistics** for θ.

- In particular, this turns out to be the case when the distribution of the complete data belongs to the **exponential family**.

- But what is a sufficient statistic? And what do we mean by a distribution belonging to the exponential family?

EM algorithm in exponential families

Sufficient statistic

↪ A **statistic** is simply a function $T = T(Y_1, \ldots, Y_n)$ of the random sample, e.g.,

$$T = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

$$T = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y}),$$

$$T = \max\{Y_1, \ldots, Y_n\}.$$

Definition

A statistic $T = T(Y_1, \ldots, Y_n)$ is said to be **sufficient** for θ if the conditional distribution of $Y_1, \ldots, Y_n$ given $T = t$ does not depend on θ for any value of t.

EM algorithm in exponential families

Sufficient statistic

- A sufficient statistic for $\boldsymbol{\theta}$ contains all the information in the sample about $\boldsymbol{\theta}$.
- Thus, given the value of T, we cannot improve our knowledge about $\boldsymbol{\theta}$ by a more detailed analysis of the data $Y_1, \ldots, Y_n$.
- This basically, and informally, means that the statistician who knows the value of T can do as just as a good job of estimating the unknown parameter $\boldsymbol{\theta}$ as the statistician who knowns the entire random sample.
- In other words, an estimate based on $T = t$ cannot be improved by using the data $Y_1, \ldots, Y_n$.

EM algorithm in exponential families
Sufficient statistic

- As an example, consider a sequence of independent Bernoulli trials

$$Y_i \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta), \qquad i = 1, \ldots, n.$$

- The number of successes $T = \sum_{i_1}^{n} Y_i$ is a sufficient statistic for the parameter θ.

- Additional information about the observed values $Y_1, \ldots, Y_n$, as e.g. the order in which the successes occurred, does not convey any information about θ.

EM algorithm in exponential families

Exponential family

Definition

One-parameter members of the exponential family have density function of the form

$$f(y;\theta) = b(y)\exp\{c(\theta)t(y) - a(\theta)\}.$$

- $\hookrightarrow$ It can be shown (see, e.g., Rice, 2006, Chapter 8) that $t(y)$ is a sufficient statistic for θ.

- $\hookrightarrow$ Suppose that $Y_1, \ldots, Y_n$ is a sample from a member of the (one-parameter) exponential family, then the joint density function is given by

$$\begin{aligned} f(\mathbf{y};\boldsymbol{\theta}) &= \prod_{i=1}^{n}\{b(y_i)\exp\{c(\theta)t(y_i) - a(\theta)\}\} \\ &= \left\{\prod_{i=1}^{n} b(y_i)\right\}\exp\left\{c(\theta)\sum_{i=1}^{n} t(y_i) - na(\theta)\right\} \end{aligned}$$

EM algorithm in exponential families

Exponential family

Definition

A p-parameter member of the exponential family has a density function of the form

$$f(y;\boldsymbol{\theta}) = b(y)\exp\left\{\sum_{j=1}^{p} c_j(\boldsymbol{\theta})t_j(y) - a(\boldsymbol{\theta})\right\}$$

- Analogously to the single parameter case, it can be shown that $(T_1(y), \ldots, T_p(y))$ is sufficient for $\boldsymbol{\theta}$.
- Suppose that $Y_1, \ldots, Y_n$ is a sample from a member of the (p-parameter) exponential family, then the joint density function is given by

$$\begin{aligned} f(\mathbf{y};\boldsymbol{\theta}) &= \prod_{i=1}^{n}\left\{b(y)\exp\left\{\sum_{j=1}^{p} c_j(\boldsymbol{\theta})t_j(y) - a(\boldsymbol{\theta})\right\}\right\} \\ &= \left\{\prod_{i=1}^{n} b(y_i)\right\}\exp\left\{\sum_{i=1}^{n}\sum_{j=1}^{p} c_j(\boldsymbol{\theta})t_j(y_i) - na(\boldsymbol{\theta})\right\} \end{aligned}$$

EM algorithm in exponential families

Exponential family

↪ Binomial distribution: $Y \sim \text{Bin}(n, \theta)$.

$$f(y) = \binom{n}{y}\theta^{y}(1-\theta)^{n-y}$$
$$= \binom{n}{y}\exp\left\{y\log\left(\frac{\theta}{1-\theta}\right) + n\log(1-\theta)\right\},$$

where

$$c(\theta) = log\left(\frac{\theta}{1-\theta}\right),$$
$$t(y) = y.$$

- Normal distribution: $Y \sim \mathrm{N}(\mu, \sigma^2)$.

$$
\begin{aligned}
f(y) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\} \\
&= \exp\left\{-\frac{1}{2\sigma^2}y^2 + \frac{\mu}{\sigma^2}y - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\},
\end{aligned}
$$

with

$$
\begin{aligned}
c(\mu, \sigma^2) &= \left(-\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2}\right), \\
t(y) &= (y^2, y).
\end{aligned}
$$

- If $Y_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$, for $i = 1, \ldots, n$, then

$$
t(\boldsymbol{y}) = \left(\sum_{i=1}^{n} y_i^2, \sum_{i=1}^{n} y_i\right).
$$

EM algorithm in exponential families

- In what follows I will be focusing in the case of a scalar parameter, but everything follows analogously to the multiparameter case.

- Let us assume that the distribution of the complete data $\mathbf{y}$ belongs to the exponential family. Then the log likelihood of the complete data can be written as

$$\log f(\mathbf{y}; \theta) = \log b(\mathbf{y}) + c(\theta)t(\mathbf{y}) - a(\theta).$$

- The E-step of the EM algorithm is

$$\begin{aligned} Q\left(\theta \mid \theta^{(t)}\right) &= E_{Y_{\text{mis}}}\left[\log b(\mathbf{y}) + c(\theta)t(\mathbf{y}) - a(\theta) \mid \mathbf{y}_{\text{obs}}, \theta^{(t)}\right] \\ &= \text{constant} + c(\theta)E_{Y_{\text{mis}}}\left[t(\mathbf{y}) \mid \mathbf{y}_{\text{obs}}, \theta^{(t)}\right] - a(\theta). \end{aligned}$$

EM algorithm in exponential families

- The M-step is

$$\frac{\mathrm{d}}{\mathrm{d}\theta}Q\left(\theta \mid \theta^{(t)}\right) = 0 \Rightarrow \frac{\mathrm{d}}{\mathrm{d}\theta}c(\theta)E_{Y_{\mathrm{mis}}}\left[t(\mathbf{y}) \mid \mathbf{y}_{\mathrm{obs}}, \theta^{(t)}\right] = \frac{\mathrm{d}}{\mathrm{d}\theta}a(\theta).$$

- From the log likelihood of the complete data, we know that

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(\mathbf{y};\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta}c(\theta)t(\mathbf{y}) - \frac{\mathrm{d}}{\mathrm{d}\theta}a(\theta).$$

- A known result from likelihood theory is that the (unconditional) expected value of the score function (derivative of the log likelihood) is zero. Therefore,

$$E\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(\mathbf{y};\theta)\right] = 0 = \frac{\mathrm{d}}{\mathrm{d}\theta}c(\theta)E[t(\mathbf{y})] - \frac{\mathrm{d}}{\mathrm{d}\theta}a(\theta) \Rightarrow \frac{\mathrm{d}}{\mathrm{d}\theta}a(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta}c(\theta)E[t(\mathbf{y})].$$

- Therefore, the M-step reduces to finding $\theta^{(t+1)}$ as a solution to

$$E_{Y_{\mathrm{mis}}}\left[t(\mathbf{y}) \mid \mathbf{y}_{\mathrm{obs}}, \theta^{(t)}\right] = E[t(\mathbf{y})].$$

EM algorithm in exponential families

Summary of the E and M steps

1. **E-step**: Compute the expected value of the sufficient statistics for the complete data, given the observed data and using the current parameter estimate $\boldsymbol{\theta}^{(t)}$. Let $\mathbf{t}^{(t)} = E_{Y_{\text{mis}}}\left[t(\mathbf{y}) \mid \mathbf{y}_{\text{obs}}, \theta^{(t)}\right]$.

2. **M-step**: Set $\boldsymbol{\theta}^{(t+1)}$ to the value that makes the unconditional expectation of the sufficient statistics for the complete data equal to $t^{(t)} = E_{Y_{\text{mis}}}\left[t(\mathbf{y}) \mid \mathbf{y}_{\text{obs}}, \theta^{(t)}\right]$. In other words, $\boldsymbol{\theta}^{(t+1)}$ solves $E[t(\mathbf{y})] = \mathbf{t}^{(t)}$.

3. Return to the E step unless a convergence criterion has been met.

EM algorithm in exponential families

Example: bivariate normal data with one variable subject to missingness

- Let us revisit the bivariate normal example from week 6 where Y_1 was fully observed but only the first m values of Y_2 were available.
- The joint density of Y_1 and Y_2 can be equivalently written as

$$f(y_1, y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{(y_1-\mu_1)^2}{\sigma_1^2} + \frac{(y_2-\mu_2)^2}{\sigma_2^2} - 2\rho\frac{(y_1-\mu_1)(y_2-\mu_2)}{\sigma_1\sigma_2}\right)\right\}.$$

- The bivariate normal distribution belongs to the exponential family.
- The sufficient statistics are

$$\left(\sum_{i=1}^{n} y_{1i}, \sum_{i=1}^{n} y_{1i}^2, \sum_{i=1}^{n} y_{2i}, \sum_{i=1}^{n} y_{2i}^2, \sum_{i=1}^{n} y_{1i}y_{2i}\right).$$

- Remember that $\theta = (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \sigma_{12})$, $\mathbf{y}_{\text{obs}} = (y_{11}, \ldots, y_{1n}, y_{21}, \ldots, y_{2m})$ and $\mathbf{y}_{\text{mis}} = (y_{2,m+1}, \ldots, y_{2n})$.
- The E-step reduces to computing

$$t_1^{(t)} = E_{Y_{\text{mis}}}\left(\sum_{i=1}^{n} y_{1i} \mid \mathbf{y}_{\text{obs}}, \theta^{(t)}\right) = \sum_{i=1}^{n} y_{1i},$$
$$t_{11}^{(t)} = E_{Y_{\text{mis}}}\left(\sum_{i=1}^{n} y_{1i}^2 \mid \mathbf{y}_{\text{obs}}, \theta^{(t)}\right) = \sum_{i=1}^{n} y_{1i}^2.$$

Note that because we do not have missing values on Y_1, the conditional expectations equal the observed values.

EM algorithm in exponential families

Example: bivariate normal data with one variable subject to missingness

$\hookrightarrow$ Because Y_2 has some missing values and the expectation of those is conditional on the current parameter estimate and on what is observed (and so, in particular, on the corresponding Y_1 values) and remembering that

$$
\begin{aligned}
&Y_2 \mid Y_1 = y_1 \sim \mathrm{N}(\beta_0 + \beta_1 y_1, \sigma^2_{2|1}),\\
&\beta_0 = \mu_2 - \beta_1 \mu_1,\\
&\beta_1 = \frac{\sigma_{12}}{\sigma^2_1},\\
&\sigma^2_{2|1} = \sigma^2_2 - \frac{\sigma^2_{12}}{\sigma^2_1}.
\end{aligned}
$$

we thus have

EM algorithm in exponential families

Example: bivariate normal data with one variable subject to missingness

$$\mathbb{E}_{y_{mis}}\left(\sum_{i=1}^{n} y_{2i} \,\middle|\, y_{obs}, \theta^{(t)}\right)$$

$$= \sum_{i=1}^{m} y_{2i} + \mathbb{E}_{y_{mis}}\left(\sum_{i=m+1}^{n} y_{2i} \,\middle|\, y_{obs}, \theta^{(t)}\right)$$

$$= \sum_{i=1}^{m} y_{2i} + \sum_{i=m+1}^{n} \mathbb{E}\left[Y_{2i} \mid y_{obs}, \theta^{(t)}\right]$$

$$= \sum_{i=1}^{m} y_{2i} + \sum_{i=m+1}^{n} \left(\beta_0^{(t)} + \beta_1^{(t)} y_{1i}\right)$$

EM algorithm in exponential families

Example: bivariate normal data with one variable subject to missingness

$$\mathbb{E}_{Y_{mis}}\left(\sum_{i=1}^{n} y_{i2}^2 \,\middle|\, y_{obs}, \theta^{(t)}\right)$$

$$= \sum_{i=1}^{m} y_{i2}^2 + \sum_{i=m+1}^{n} \mathbb{E}\left[Y_{2i}^2 \,\middle|\, y_{obs}, \theta^{(t)}\right]$$

$$= \sum_{i=1}^{m} y_{i2}^2 + \sum_{i=m+1}^{n} \left(\left(\sigma_{2|1}^{(t)}\right)^2 + \left(\beta_0^{(t)} + \beta_1^{(t)} y_{1i}\right)^2 \right)$$

EM algorithm in exponential families

Example: bivariate normal data with one variable subject to missingness

$$\mathbb{E}_{Y_{mis}}\left(\sum_{i=1}^{n} y_{1i} y_{2i} \,\Big|\, y_{obs}, \theta^{(t)} \right)$$

$$= \sum_{i=1}^{m} y_{1i} y_{2i} + \sum_{i=m+1}^{n} \mathbb{E}_{y_{mis}} \left[\underline{y_{1i} y_{2i}} \,\Big|\, y_{obs}, \theta^{(t)} \right]$$

$$= \sum_{i=1}^{m} y_{1i} y_{2i} + \sum_{i=m+1}^{n} y_{1i} \mathbb{E}\left[Y_{2i} \,\Big|\, y_{obs}, \theta^{(t)} \right]$$

$$= \sum_{i=1}^{m} y_{1i} y_{2i} + \sum_{i=m+1}^{n} y_{1i} \left(\beta_0^{(t)} + \beta_1^{(t)} y_{1i} \right)$$

EM algorithm in exponential families

Example: bivariate normal data with one variable subject to missingness

EM algorithm in exponential families

Example: bivariate normal data with one variable subject to missingness

EM algorithm in exponential families

Example: bivariate normal data with one variable subject to missingness

$\hookrightarrow$ For the M-step we need

$$t_1^{(t)} = E\left(\sum_{i=1}^{n} y_{1i}\right) = n\mu_1,$$

$$t_{11}^{(t)} = E\left(\sum_{i=1}^{n} y_{1i}^2\right) = n(\mu_1^2 + \sigma_1^2),$$

$$t_2^{(t)} = E\left(\sum_{i=1}^{n} y_{2i}\right) = n\mu_2,$$

$$t_{22}^{(t)} = E\left(\sum_{i=1}^{n} y_{2i}^2\right) = n(\mu_2^2 + \sigma_2^2),$$

$$t_{12}^{(t)} = E\left(\sum_{i=1}^{n} y_{1i}y_{2i}\right) = n(\sigma_{12} + \mu_1\mu_2).$$

↪ We thus have

$$
\begin{aligned}
\mu_1^{(t+1)} &= \frac{t_1^{(t)}}{n},\\
\left(\sigma_1^{(t+1)}\right)^2 &= \frac{t_{11}^{(t)}}{n} - (\mu_1^{(t+1)})^2,\\
\mu_2^{(t+1)} &= \frac{t_2^{(t)}}{n},\\
\left(\sigma_2^{(t+1)}\right)^2 &= \frac{t_{22}^{(t)}}{n} - (\mu_2^{(t+1)})^2,\\
\sigma_{12}^{(t+1)} &= \frac{t_{12}^{(t)}}{n} - \mu_1^{(t+1)}\mu_2^{(t+1)}.
\end{aligned}
$$

↪ Note that $t_1^{(t)}$, $t_{11}^{(t)}$, $\mu_1^{(t+1)}$, and $\sigma_1^{(t+1)}$ are constant across iterations because there are no missing values in Y_1.