Incomplete Data Analysis

Vanda Inácio

University of Edinburgh

Semester 1, 2018/2019

EM algorithm

ABO blood type

- The locus corresponding to the ABO blood group has three alleles A, B, and O, and is located on chromosome 9q34. Alleles A and B are co-dominant and the alleles A and B are dominant to O.

- This leads to the following genotypes (unobserved) and phenotypes/blood groups (observed)

Genotype	AA	AO	AB	BB	BO	OO
Phenotype	A	A	AB	B	B	O

- Accordingly to the Hardy–Weinberg equilibrium law (the types of the two alleles carried by an individual are independent), the probabilities associated to each genotype are

Genotype	AA	AO	AB	BB	BO	OO
Probability	p_A^2	$2p_Ap_O$	$2p_Ap_B$	p_B^2	$2p_Bp_O$	p_O^2

where p_A, p_B, and p_O are the probabilities of occurrence of alleles A, B, and O, respectively.

EM algorithm

ABO blood type

- From a sample of 521 individuals, the following blood types were observed

Blood type	A	B	AB	O
Total number	186	38	13	284

- Our goal is to estimate p_A, p_B, and p_O. Let $p = (p_A, p_B, p_O)$ and note that $p_A + p_B + p_O = 1$ (and so, for instance, $p_O = 1 - p_A - p_B$).

- What is the observed data? What is the complete data?

- The complete data is $y = (n_{AA}, n_{AO}, n_{AB}, n_{BB}, n_{BO}, n_{OO})$, where n_{AA} is the number of people with genotype AA, n_{AO} is the number of people with genotype AO, etc.

- The observed data is $y_{\text{obs}} = (n_A, n_B, n_{AB}, n_O)$, where n_A is the number of people with blood type A, n_B is the number of people with blood type B, etc.

- Note that $n_{AA} + n_{AO} = n_A$, $n_{BB} + n_{BO} = n_B$, and $n_{OO} = n_O$.

EM algorithm

ABO blood type

- Let n be the total sample size.
- The likelihood of the complete data has the following multinomial form

$$L(p_A, p_B; y) = \frac{n!}{n_{AA}!n_{AO}!n_{BB}!n_{BO}!n_{AB}!n_{OO}!}(p_A^2)^{n_{AA}}(2p_Ap_O)^{n_{AO}}(p_B^2)^{n_{BB}}(2p_Bp_O)^{n_{BO}}(2p_Ap_B)^{n_{AB}}(p_O^2)^{n_{OO}}$$

- The complete data loglikelihood is then proportional to

$$\begin{aligned}\log L(p_A, p_B; y) &\propto 2n_{AA}\log p_A + n_{AO}(\log 2 + \log p_A + \log p_O) + 2n_{BB}\log p_B \\ &\quad + n_{BO}(\log 2 + \log p_B + \log p_O) + n_{AB}(\log 2 + \log p_A + \log p_B) + 2n_{OO}\log p_O \\ &\propto (2n_{AA} + n_{AO} + n_{AB})\log p_A + (2n_{BB} + n_{BO} + n_{AB})\log p_B + (n_{AO} + n_{BO} + 2n_{OO})\log p_O\end{aligned}$$

- Here, both n_{AA}, n_{AO}, n_{BB}, and n_{BO} are unknown. Let us replace n_{AO} by $n_A - n_{AA}$ and n_{BO} by $n_B - n_{BB}$. Now, only n_{AA} and n_{BB} are unknown. We will replace also p_O by $1 - p_A - p_B$ and n_{OO} by n_O.

EM algorithm

ABO blood type

↪ Then,

$$\log L(p_A, p_B; y) \propto (n_{AA} + n_A + n_{AB}) \log p_A + (n_{BB} + n_B + n_{AB}) \log p_B + (2n_O + n_A + n_B - n_{AA} - n_{BB}) \log(1 - p_A - p_B).$$

↪ The Q function from the E-step is given by

$$\begin{aligned} Q(p \mid p^{(t)}) &= E_{N_{AA}, N_{BB}} \left[\log L(p_A, p_B; y) \mid n_A, n_B, n_{AB}, n_O, p^{(t)} \right] \\ &= (E[N_{AA} \mid n_A, p^{(t)}] + n_A + n_{AB}) \log p_A + (E[N_{BB} \mid n_B, p^{(t)}] + n_B + n_{AB}) \log p_B \\ &+ (2n_O + n_A + n_B - E[N_{AA} \mid n_A, p^{(t)}] - E[N_{BB} \mid n_B, p^{(t)}]) \log(1 - p_A - p_B). \end{aligned}$$

EM algorithm

ABO blood type

↪ Note that

$$N_{AA} \mid N_A = n_A \sim \text{Bin}(n_A, p_1), \qquad p_1 = \frac{p_A^2}{p_A^2 + 2p_A p_O} = \frac{p_A}{2 - p_A - 2p_B}.$$

↪ Similarly,

$$N_{BB} \mid N_B = n_B \sim \text{Bin}(n_B, p_2), \qquad p_2 = \frac{p_B^2}{p_B^2 + 2p_B p_O} = \frac{p_B}{2 - p_B - 2p_A}.$$

↪ Thus,

$$E[N_{AA} \mid n_A, p^{(t)}] = n_A \frac{p_A^{(t)}}{2 - p_A^{(t)} - 2p_B^{(t)}} = n_{AA}^{(t)},$$

$$E[N_{BB} \mid n_B, p^{(t)}] = n_B \frac{p_B^{(t)}}{2 - p_B^{(t)} - 2p_A^{(t)}} = n_{BB}^{(t)}.$$

EM algorithm

ABO blood type

↪ We complete the E-step as

$$Q(p \mid p^{(t)}) = (n_{AA}^{(t)} + n_A + n_{AB}) \log p_A + (n_{BB}^{(t)} + n_B + n_{AB}) \log p_B + (2n_O + n_A + n_B - n_{AA}^{(t)} - n_{BB}^{(t)}) \log(1 - p_A - p_B).$$

↪ For the M-step, we need

$$\frac{\partial}{\partial p_A} Q(p \mid p^{(t)}) = 0 \Rightarrow \frac{1}{p_A}(n_{AA}^{(t)} + n_A + n_{AB}) - \frac{1}{1 - p_A - p_B}(2n_O + n_A + n_B - n_{AA}^{(t)} - n_{BB}^{(t)}) = 0,$$

$$\frac{\partial}{\partial p_B} Q(p \mid p^{(t)}) = 0 \Rightarrow \frac{1}{p_B}(n_{BB}^{(t)} + n_B + n_{AB}) - \frac{1}{1 - p_A - p_B}(2n_O + n_A + n_B - n_{AA}^{(t)} - n_{BB}^{(t)}) = 0.$$

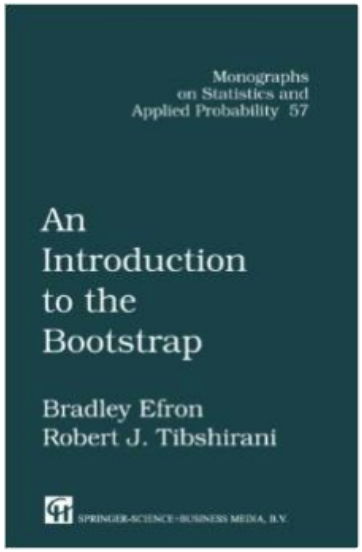
↪ Solving the above system of equations, we obtain

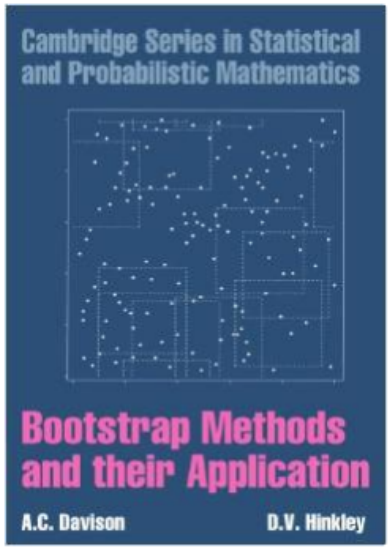
$$p_A^{(t+1)} = \frac{n_{AA}^{(t)} + n_A + n_{AB}}{2n}, \qquad p_B^{(t+1)} = \frac{n_{BB}^{(t)} + n_B + n_{AB}}{2n}.$$

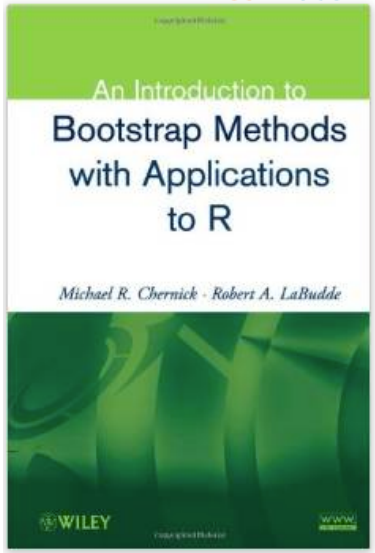
EM algorithm

Bootstrap and the EM (not examinable)

- So far we have been applying the EM algorithm to obtain maximum likelihood estimates for a broad type of incomplete data problems.
- A disadvantage is that the standard error estimates are not readily available at the conclusion of the algorithm as they would be, for example, from direct maximisation of the observed data likelihood using optimisation techniques, such as the Newton–Raphson algorithm, in which the observed information matrix is calculated at each iteration.
- We discuss a general approach to obtain standard error estimates/confidence intervals: the bootstrap (Effron, 1979).







EM algorithm

Bootstrap and the EM (not examinable)

- The essential concept of bootstrapping is to emulate repetition of the experiment by simulating new data on the computer, followed by recalculation of the mle using the simulated data.
- The bootstrap effectively replaces the calculus and theory (to obtain standard errors/confidence intervals) with pure computational effort. But, of course, does not eliminate the need of thinking!
- The term ‘bootstrap’ was chosen by Efron (1979) in the first comprehensive account of this computer-intensive methodology.

EM algorithm

Bootstrap and the EM (not examinable)

- ↪ The bootstrap emulates the sampling distribution of $\widehat{\theta}$ by emulating the processes of data generation and model fitting.
- ↪ It does this, by generating artificial data, say $y^{(b)} = (y_1^{(b)}, \ldots, y_n^{(b)})$ from a distribution that approximates the true unknown sampling distribution of the actual data, followed by recalculating the mle using the artificial data.
- ↪ This is done a large number of times, say B, resulting in a large collection of bootstrap mle's, denoted $\widehat{\theta}^{(b)}$, $b = 1, \ldots, B$.
- ↪ The distribution of these artificially generated bootstrap mle's can then be used to infer the sampling distribution of $\widehat{\theta}$.

EM algorithm

Bootstrap and the EM (not examinable)

- One obvious choice for a distribution to approximate the true unknown sampling distribution of the data is to use the distribution under the fitted model.
- That is, to generate $y^{(b)}$ from the distribution with density $f(.;\widehat{\theta})$. This is parametric bootstrapping.
- Nonparametric bootstrapping is an alternative to the parametric bootstrap.
- If the data y_i, $i = 1, \ldots, n$ are iid then the empirical cumulative distribution function (ecdf)

$$\widehat{F}_{ecdf}(y) = \frac{1}{n}\sum_{i=1}^{n} I(y_i \leq y)$$

can be used as a discrete approximation to the true unknown cumulative distribution function.

EM algorithm

Bootstrap and the EM (not examinable)

↪ Note that the ecdf assigns probability mass $1/n$ to each y_i value (and if two or more data points take the same value then the probability mass for that value is summed over those data points).

↪ The nonparametric bootstrap generates new data $y^{(b)}$ by random sampling from the ecdf. Note that random sampling from the ecdf can be accomplished by sampling with replacement from the observed data $y_1, \ldots, y_n$.

EM algorithm

Bootstrap and the EM (not examinable)

↪ The general algorithm is as follows. For $b = 1, \ldots, B$

1. Generate a bootstrap sample $y^{(b)} = (y_1^{(b)}, \ldots, y_n^{(b)}) \sim \widehat{F}$ ($\widehat{F}$ can be parametric or nonparametric).
2. Compute $\widehat{\theta}^{(b)}$ using $y^{(b)} = (y_1^{(b)}, \ldots, y_n^{(b)})$.

↪ The bootstrap variance can be calculated as

$$\text{var}_{\text{boot}}(\widehat{\theta}) = \frac{1}{B-1} \sum_{b=1}^{B} (\widehat{\theta}^{(b)} - \bar{\widehat{\theta}}^{*})^2,$$

where $\bar{\widehat{\theta}}^{*} = \frac{1}{B} \sum_{b=1}^{B} \widehat{\theta}^{(b)}$.

EM algorithm

Bootstrap and the EM (not examinable)

- There are several ways to compute bootstrap confidence intervals, we restrict ourselves to the most commonly used form of bootstrap confidence interval, the percentile method.
- This method is the simplest and it performs well in most situations.
- Under this method, a $100(1-\alpha)\%$ bootstrap confidence interval is computed as

$$(\widehat{\theta}^{(b,l)}, \widehat{\theta}^{(b,u)}),$$

where $\widehat{\theta}^{(b,l)}$ and $\widehat{\theta}^{(b,u)}$ are the empirical $\alpha/2$ and $1-\alpha/2$ quantiles of the collection

$$(\widehat{\theta}^{(1)}, \ldots, \widehat{\theta}^{(B)}).$$

EM algorithm

Bootstrap and the EM (not examinable)

- How many bootstrap simulations is enough?
- The number of bootstrap simulations B must be large enough that the estimated standard errors or quantiles are not materially affected by the inherent randomness of the bootstrap.
- That is, they should be negligibly if the bootstrap were to be repeated.
- Davison and Hinkley (1997) recommended performing at least 1000 bootstrap simulations when calculating 95% bootstrap confidence intervals.

EM algorithm

Bootstrap and the EM (not examinable)

↪ See the supplementary file (available on Learn) for an application of the parametric and nonparametric bootstrap to the linkage model example.

↪ On such file is also available an application of the nonparametric bootstrap to the old faithful dataset/two component mixture. You can try to implement the parametric bootstrap by your own (it is easy)!

EM algorithm

Bootstrap and the EM (not examinable)

↪ Finally, for those interested in learning more about the EM algorithm, I leave the following reference.