3.3 Problems with incomplete observations

3.3.1 Expectation-maximization algorithm

Maximizing likelihood functions receives special attention in statistics and optimization theory, because likelihood functions often have special forms. Many optimization algorithms aim at approximating the target function locally with an easy-to-maximize surrogate function. Then one optimizes the surrogate function iteratively until the algorithm converges to a local optimum. Newton-Raphson algorithm is one example of such an iterative optimization strategy, where the surrogate function is taken to be a quadratic function. Expectation-maximization algorithm provides a different strategy for statistical problems with incomplete observations (Dempster et al., 1977).

We start with the complete data $\mathbf{x}$ and **complete data likelihood**

$$\Pr(\mathbf{x};\boldsymbol{\theta}).$$

In order to avoid measure-theoretic difficulties, we will consider only discrete $\mathbf{x}$. Suppose we observe only some transformation of the complete data: $\mathbf{y} = h(\mathbf{x})$. Then the **observed data likelihood**, the probability of observing $\mathbf{y}$, is computed by summing over all values of $\mathbf{x}$, compatible with $\mathbf{y}$:

$$\Pr(\mathbf{y};\boldsymbol{\theta}) = \sum_{\mathbf{x}:\mathbf{y}=h(\mathbf{x})} \Pr(\mathbf{x};\boldsymbol{\theta}).$$

The following algorithm iteratively maximizes the observed data likelihood.

Algorithm 6 EM algorithm for maximizing $\Pr(\mathbf{y};\boldsymbol{\theta})$

Start with an arbitrary $\boldsymbol{\theta}_0$
repeat
  Set $\boldsymbol{\theta}_{n+1} = \arg\max_{\boldsymbol{\theta}} \mathrm{E}\left[\ln \Pr(\mathbf{x};\boldsymbol{\theta}) \mid \mathbf{y};\boldsymbol{\theta}_n\right]$
until $||\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n||_\infty < \varepsilon$ (another possibility: $|\ln \Pr(\mathbf{y};\boldsymbol{\theta}_{n+1}) - \ln \Pr(\mathbf{y};\boldsymbol{\theta}_n)| < \varepsilon$)
return $\boldsymbol{\theta}_{n+1}$ and $\Pr(\mathbf{y};\boldsymbol{\theta}_{n+1})$

Example: *ABO blood types* Suppose we sample n individuals from a population and for each individual we record his blood type, $f_i \in \{A, B, AB, O\}$. We recall that blood types have underlying genotypes:

$$\begin{aligned}
g_i = AA \text{ or } AO &\Rightarrow f_i = A \\
g_i = BB \text{ or } BO &\Rightarrow f_i = B \\
g_i = AB &\Rightarrow f_i = AB \\
g_i = OO &\Rightarrow f_i = OO.
\end{aligned}$$

Therefore, $\mathbf{g} = (g_1, \ldots, g_n)$ is the complete data and $\mathbf{f} = (f_1, \ldots, f_n)$ is the observed data. According to the Hardy-Weinberg equilibrium, the complete data likelihood is

$$\Pr(\mathbf{g};\mathbf{p}) \propto (p_A^2)^{m_{AA}} (2p_A p_O)^{m_{AO}} (p_B^2)^{m_{BB}} (2p_B p_O)^{m_{BO}} (2p_A p_B)^{m_{AB}} (p_O^2)^{m_{OO}},$$

where $\mathbf{p} = (p_A, p_B, p_O)$ are allele frequencies, and

$$\begin{aligned}
m_{AA} &= \sum_{i=1}^{n} 1_{\{g_i = AA\}} \qquad m_{AO} = \sum_{i=1}^{n} 1_{\{g_i = AO\}} \qquad m_{BB} = \sum_{i=1}^{n} 1_{\{g_i = BB\}} \\
m_{BO} &= \sum_{i=1}^{n} 1_{\{g_i = BO\}} \qquad m_{AB} = \sum_{i=1}^{n} 1_{\{g_i = AB\}} \qquad m_{OO} = \sum_{i=1}^{n} 1_{\{g_i = OO\}}.
\end{aligned}$$

The observed data likelihood becomes

$$\Pr(\mathbf{f};\mathbf{p}) \propto (p_A^2+2p_Ap_O)^{n_A}(p_B^2+2p_Bp_O)^{n_B}(2p_Ap_B)^{n_{AB}}(p_O^2)^{n_O},$$

where

$$n_A=\sum_{i=1}^{n}1_{\{f_i=A\}} \qquad n_B=\sum_{i=1}^{n}1_{\{f_i=B\}}$$
$$n_{AB}=\sum_{i=1}^{n}1_{\{f_i=AB\}} \qquad m_O=\sum_{i=1}^{n}1_{\{f_i=O\}}.$$

The observed data likelihood equations, obtained by differentiating the log likelihood and setting these derivatives to 0, are non-linear with no closed-form solution. To set up an EM algorithm, we first consider the expected log complete data likelihood, where expectations is taken using the model parameters from the kth iteration:

$$\begin{aligned}&\mathrm{E}\left[\ln\Pr(\mathbf{g};\mathbf{p})\,|\,\mathbf{f};\mathbf{p}_k\right]=2\mathrm{E}\left(m_{AA}\,|\,\mathbf{f};\mathbf{p}_k\right)\ln p_A+2\mathrm{E}\left(m_{BB}\,|\,\mathbf{f};\mathbf{p}_k\right)\ln p_B+\mathrm{E}\left(m_{AO}\,|\,\mathbf{f};\mathbf{p}_k\right)\left(\ln p_A+\ln p_O\right)\\&+\mathrm{E}\left(m_{BO}\,|\,\mathbf{f};\mathbf{p}_k\right)\left(\ln p_B+\ln p_O\right)+\mathrm{E}\left(m_{AB}\,|\,\mathbf{f};\mathbf{p}_k\right)\left(\ln p_A+\ln p_B\right)+2\mathrm{E}\left(m_{OO}\,|\,\mathbf{f};\mathbf{p}_k\right)\ln p_O+C.\end{aligned}$$

$$\begin{aligned}&m_{k,AA}=\mathrm{E}\left(m_{AA}\,|\,\mathbf{f};\mathbf{p}_k\right)=\mathrm{E}\left(\sum_{i=1}^{n}1_{\{g_i=AA\}}\,|\,\mathbf{f};\mathbf{p}_k\right)=\sum_{i=1}^{n}\mathrm{E}\left(1_{\{g_i=AA\}}\,|\,\mathbf{f};\mathbf{p}_k\right)\\&=\sum_{i=1}^{n}\mathrm{E}\left(1_{\{g_i=AA\}}\,|\,f_i;\mathbf{p}_k\right)=\sum_{i=1}^{n}1_{\{f_i=A\}}\Pr\left(g_i=AA\,|\,f_i=A;\mathbf{p}_k\right)=n_A\Pr\left(g_1=AA\,|\,f_1=A;\mathbf{p}_k\right)\\&=n_A\frac{\Pr\left(g_1=AA,f_1=A;\mathbf{p}_k\right)}{\Pr\left(f_1=A;\mathbf{p}_k\right)}=n_A\frac{\Pr\left(g_1=AA;\mathbf{p}_k\right)}{\Pr\left(f_1=A;\mathbf{p}_k\right)}=n_A\frac{p_{k,A}^2}{p_{k,A}^2+2p_{k,A}p_{k,O}}.\end{aligned}$$

The other conditional expectations are derived in the same manner:

$$\begin{aligned}m_{k,AA}&=\mathrm{E}\left(m_{AA}\,|\,\mathbf{f};\mathbf{p}_k\right)=n_A\frac{p_{k,A}^2}{p_{k,A}^2+2p_{k,A}p_{k,O}}\\m_{k,AO}&=\mathrm{E}\left(m_{AO}\,|\,\mathbf{f};\mathbf{p}_k\right)=n_A\frac{2p_{k,A}p_{k,O}}{p_{k,A}^2+2p_{k,A}p_{k,O}}\\m_{k,BB}&=\mathrm{E}\left(m_{BB}\,|\,\mathbf{f};\mathbf{p}_k\right)=n_B\frac{p_{k,B}^2}{p_{k,B}^2+2p_{k,B}p_{k,O}}\\m_{k,BO}&=\mathrm{E}\left(m_{BO}\,|\,\mathbf{f};\mathbf{p}_k\right)=n_B\frac{2p_{k,B}p_{k,O}}{p_{k,B}^2+2p_{k,B}p_{k,O}}\\m_{k,AB}&=\mathrm{E}\left(m_{AB}\,|\,\mathbf{f};\mathbf{p}_k\right)=n_{AB}\\m_{k,OO}&=\mathrm{E}\left(m_{OO}\,|\,\mathbf{f};\mathbf{p}_k\right)=n_O.\end{aligned}$$

In the M-step of the EM algorithm, we need to maximize the expected log likelihood of the complete data:

$$\begin{aligned}&H(\mathbf{p})=\mathrm{E}\left[\ln\Pr(\mathbf{g};\mathbf{p})\,|\,\mathbf{f};\mathbf{p}_k\right]=2m_{k,AA}\ln p_A+2m_{k,BB}\ln p_B+m_{k,AO}(\ln p_A+\ln p_O)\\&+m_{k,BO}(\ln p_B+\ln p_O)+n_{AB}(\ln p_A+\ln p_B)+2n_O\ln p_O,\end{aligned}$$

Using the method of Lagrange multipliers, we form a surrogate function

$$F(\mathbf{p},\lambda)=H(\mathbf{p})+\lambda(p_A+p_B+p_o-1)$$

and set the gradient of this function to 0:

$$\begin{aligned}
\frac{\partial F}{\partial p_A} &= \frac{2m_{k,AA} + m_{k,AO} + n_{AB}}{p_A} + \lambda = 0 \\
\frac{\partial F}{\partial p_B} &= \frac{2m_{k,BB} + m_{k,BO} + n_{AB}}{p_B} + \lambda = 0 \\
\frac{\partial F}{\partial p_O} &= \frac{2n_O + m_{k,AO} + n_{AB}}{p_O} + \lambda = 0 \\
\frac{\partial F}{\partial \lambda} &= p_A + p_B + p_O - 1 = 0.
\end{aligned}$$

Solving these equations yields

$$\begin{aligned}
p_{k+1,A} &= \frac{2m_{k,AA} + m_{k,AO} + n_{AB}}{2n} \\
p_{k+1,B} &= \frac{2m_{k,BB} + m_{k,BO} + n_{AB}}{2n} \\
p_{k+1,O} &= \frac{2n_O + m_{k,AO} + m_{k,BO}}{2n}.
\end{aligned}$$

We omit the proof that this critical point is indeed a maximum.

Note 3.6. When the complete data likelihood belongs to the exponential family, i.e.

$$l(\mathbf{x};\boldsymbol{\theta}) = \ln a(\mathbf{x}) + \boldsymbol{\theta}^T \mathbf{T}(\mathbf{x}) - \ln b(\boldsymbol{\theta}),$$

the expectation step in the EM algorithm always reduces to the taking expectation of the sufficient statistics, $\mathrm{E}[\mathbf{T}(\mathbf{x}) \,|\, \mathbf{y}; \boldsymbol{\theta}_k]$

Note 3.7. The EM algorithm is not necessarily the best way to maximize the likelihood. One argument against using EM is the fact that the algorithm converges linearly to a local maximum. Newton-Raphson, for example, enjoys the quadratic rate of convergence. On the other hand, every step of the EM algorithm is guaranteed not to decrease the observed data likelihood. This is called the ascent property of the EM algorithm. To prove this result, we will need Jensen's inequality.

Definition. A function $f : \mathbb{R}^n \to \mathbb{R}$ is called **convex** if its domain is a convex set and for any $\alpha \in [0,1]$ and $x, y \in \mathbb{R}^n$

$$\alpha f(x) + (1-\alpha) f(y) \geq f[\alpha x + (1-\alpha) y]$$

Jensen's inequality:

If g is a convex function, then $\mathrm{E}[g(X)] \geq g[\mathrm{E}(X)]$.

The ascent property of EM algorithm:

$\ln \Pr(\mathbf{y}; \boldsymbol{\theta}_{n+1}) \geq \ln \Pr(\mathbf{y}; \boldsymbol{\theta}_n)$.

Proof. Assuming that

$$\boldsymbol{\theta}_{n+1} = \arg\max_{\boldsymbol{\theta}} \mathrm{E}\left[\ln \Pr(\mathbf{x};\boldsymbol{\theta}) \,|\, \mathbf{y}; \boldsymbol{\theta}_n\right], \tag{3.11}$$

we write

$$\begin{aligned}
&\ln \Pr(\mathbf{y};\boldsymbol{\theta}_{n+1}) - \ln \Pr(\mathbf{y};\boldsymbol{\theta}_n) = [\text{cond prob}] \\
&= \ln \Pr(\mathbf{x};\boldsymbol{\theta}_{n+1}) - \ln \Pr(\mathbf{x};\boldsymbol{\theta}_n) + \ln \Pr(\mathbf{x} \,|\, \mathbf{y};\boldsymbol{\theta}_n) - \ln \Pr(\mathbf{x} \,|\, \mathbf{y};\boldsymbol{\theta}_{n+1}).
\end{aligned} \tag{3.12}$$

Now, taking expectation $\mathrm{E}(\cdot\,|\,\mathbf{y};\boldsymbol{\theta}_n)$ over $\mathbf{x}$ of both sides of (3.12), we arrive at

$$\begin{aligned}
&\ln\Pr(\mathbf{y};\boldsymbol{\theta}_{n+1}) - \ln\Pr(\mathbf{y};\boldsymbol{\theta}_n) = \mathrm{E}\left(\ln\Pr(\mathbf{x};\boldsymbol{\theta}_{n+1})\,|\,\mathbf{y};\boldsymbol{\theta}_n\right) - \mathrm{E}\left(\ln\Pr(\mathbf{x};\boldsymbol{\theta}_n)\,|\,\mathbf{y};\boldsymbol{\theta}_n\right)\\
&+\mathrm{E}\left(\ln\Pr(\mathbf{x}\,|\,\mathbf{y};\boldsymbol{\theta}_n)\,|\,\mathbf{y};\boldsymbol{\theta}_n\right) - \mathrm{E}\left(\ln\Pr(\mathbf{x}\,|\,\mathbf{y};\boldsymbol{\theta}_{n+1})\,|\,\mathbf{y};\boldsymbol{\theta}_n\right) \geq [\text{condition (3.11)}]\\
&\geq +\mathrm{E}\left(\ln\Pr(\mathbf{x}\,|\,\mathbf{y};\boldsymbol{\theta}_n)\,|\,\mathbf{y};\boldsymbol{\theta}_n\right) - \mathrm{E}\left(\ln\Pr(\mathbf{x}\,|\,\mathbf{y};\boldsymbol{\theta}_{n+1})\,|\,\mathbf{y};\boldsymbol{\theta}_n\right)\\
&= \mathrm{E}\left[-\ln\left(\frac{\Pr(\mathbf{x}\,|\,\mathbf{y};\boldsymbol{\theta}_{n+1})}{\Pr(\mathbf{x}\,|\,\mathbf{y};\boldsymbol{\theta}_n)}\right)\,|\,\mathbf{y};\boldsymbol{\theta}_n\right] \geq [\text{Jensen's inequality}]\\
&\geq -\ln\mathrm{E}\left[\left(\frac{\Pr(\mathbf{x}\,|\,\mathbf{y};\boldsymbol{\theta}_{n+1})}{\Pr(\mathbf{x}\,|\,\mathbf{y};\boldsymbol{\theta}_n)}\right)\,|\,\mathbf{y};\boldsymbol{\theta}_n\right] = -\ln\left(\sum_{\mathbf{x}}\frac{\Pr(\mathbf{x}\,|\,\mathbf{y};\boldsymbol{\theta}_{n+1})}{\Pr(\mathbf{x}\,|\,\mathbf{y};\boldsymbol{\theta}_n)}\Pr(\mathbf{x}\,|\,\mathbf{y};\boldsymbol{\theta}_n)\right)\\
&= -\ln\left(\sum_{\mathbf{x}}\Pr(\mathbf{x}\,|\,\mathbf{y};\boldsymbol{\theta}_{n+1})\right) = -\ln 1 = 0.
\end{aligned}$$

□

Note 3.8. Particular cases of the EM algorithm, known as gene counting, were in use in genetics since the 1950s (Ceppelini et al., 1955). Baum et al. (1970) used the EM algorithm for fitting hidden Markov models and proved the ascent property of the EM algorithm. Finally, Sundberg (1976) introduced the EM algorithm for situations where the complete data likelihood belongs to exponential families. (Dempster et al., 1977)'s paper, which is now a classic EM algorithm reference, arrived just in time to synthesize all this knowledge, although the authors overstated the convergence properties of the algorithm.

3.3.2 Bayesian data augmentation

As before, we start with complete data $\mathbf{x}$, the **complete data likelihood**

$$\Pr(\mathbf{x}\,|\,\boldsymbol{\theta}),$$

incomplete data: $\mathbf{y} = h(\mathbf{x})$, and the **observed data likelihood**

$$\Pr(\mathbf{y}\,|\,\boldsymbol{\theta}) = \sum_{\mathbf{x}:\mathbf{y}=h(\mathbf{x})}\Pr(\mathbf{x};\boldsymbol{\theta}).$$

We assume a prior distribution of model parameters, $\Pr(\boldsymbol{\theta})$. Normally, we would proceed by approximating the posterior distribution $\Pr(\boldsymbol{\theta}\,|\,\mathbf{y})$. Instead, we augment the the parameter state space with complete data and form the augmented posterior distribution:

$$\Pr(\boldsymbol{\theta},\mathbf{x}\,|\,\mathbf{y}) \propto \Pr(\boldsymbol{\theta},\mathbf{x},\mathbf{y}) = \Pr(\boldsymbol{\theta},\mathbf{x})\,1_{\{\mathbf{y}=h(\mathbf{x})\}} \propto \Pr(\mathbf{x}\,|\,\boldsymbol{\theta})\Pr(\boldsymbol{\theta})\,1_{\{\mathbf{y}=h(\mathbf{x})\}}.$$

In many applications, it is convenient to iterate between updating complete data conditional on parameters and updating parameters conditional on complete data. In the example below, it is possible to accomplish this iterative procedure by a two-stage Gibbs sampler.

Example: *ABO blood type example (continued)* As before, the complete data likelihood is

$$\Pr(\mathbf{g}\,|\,\mathbf{p}) \propto (p_A^2)^{m_{AA}}(2p_Ap_O)^{m_{AO}}(p_B^2)^{m_{BB}}(2p_Bp_O)^{m_{BO}}(2p_Ap_B)^{m_{AB}}(p_O^2)^{m_{OO}}.$$

We put Dirichlet$(\gamma_A,\gamma_B,\gamma_O)$ prior on the allele frequencies:

$$\Pr(p_A,p_B,p_O) \propto p_A^{\gamma_A-1}p_B^{\gamma_B-1}p_O^{\gamma_O-1}.$$

Sampling from the posterior

$$\Pr(\mathbf{p}\,|\,\mathbf{f}) \propto \Pr(\mathbf{f}\,|\,\mathbf{p})\Pr(\mathbf{p})$$

is possible, but not pretty. Instead let us augment our model parameters with missing data. The augmented posterior distribution becomes

$$\Pr(\mathbf{p},\mathbf{g}\,|\,\mathbf{f}) \propto \Pr(\mathbf{g}\,|\,\mathbf{p})\Pr(\mathbf{p}),$$

where $\mathbf{g}$ in the complete data likelihood must be compatible with observed $\mathbf{f}$. Conditional on complete data, the density of $\mathbf{p}$ is

$$\Pr(\mathbf{p}\,|\,\mathbf{g},\mathbf{f}) \propto \Pr(\mathbf{g}\,|\,\mathbf{p})\Pr(\mathbf{p}) \propto p_A^{2m_{AA}+m_{AO}+m_{AB}+\gamma_A-1} p_B^{2m_{BB}+m_{BO}+m_{AB}+\gamma_B-1} p_O^{2m_{OO}+m_{BO}+m_{AO}+\gamma_O-1}.$$

Therefore,

$$\mathbf{p} \mid \mathbf{g},\mathbf{f} \sim \text{Dirichlet}(2m_{AA}+m_{AO}+m_{AB}+\gamma_A, 2m_{BB}+m_{BO}+m_{AB}+\gamma_B, 2m_{OO}+m_{BO}+m_{AO}+\gamma_O).$$

Next, we want to understand the distribution of complete data, conditional on the allele frequencies and observed data, $\Pr(\mathbf{g}\,|\,\mathbf{p},\mathbf{f})$. It is convenient to start thinking about unobserved genotypes of individual i. If $f_i = AB$ or $f_i = O$, then there is no ambiguity leading $\Pr(g_i = AB\,|\,f_i = AB) = 1$ and $\Pr(g_i = OO\,|\,f_i = O) = 1$. For the other two phenotypes, we have

$$\Pr(g_i = AA\,|\,f_i = A,\mathbf{p}) = \frac{p_A^2}{p_A^2+2p_Ap_O}, \qquad \Pr(g_i = AO\,|\,f_i = A,\mathbf{p}) = \frac{2p_Ap_O}{p_A^2+2p_Ap_O},$$
$$\Pr(g_i = BB\,|\,f_i = A,\mathbf{p}) = \frac{p_B^2}{p_B^2+2p_Bp_O}, \qquad \Pr(g_i = BO\,|\,f_i = A,\mathbf{p}) = \frac{2p_Bp_O}{p_B^2+2p_Bp_O}.$$

Since in this example, we are not interested in imputing individual genotypes and individual genotypes are independent give allele frequencies, we can sample sufficient statistics of the complete data directly:

$$m_{AA} \mid \mathbf{f},\mathbf{p} \sim \text{Bin}\left(n_A, \frac{p_A}{p_A+2p_O}\right) (\text{set } m_{AO} = n_A - m_{AA}),$$
$$m_{BB} \mid \mathbf{f},\mathbf{p} \sim \text{Bin}\left(n_B, \frac{p_B}{p_B+2p_O}\right) (\text{set } m_{BO} = n_B - m_{BB}).$$

Alternating between sampling $\mathbf{p} \mid \mathbf{m}$ and $\mathbf{m} \mid \mathbf{p},\mathbf{n}$ we form a Gibbs sampling algorithm on the state space of model parameters and complete data sufficient statistics.

Note 3.9. Notice the similarities between the EM algorithm and Bayesian data augmentation. In both cases, we impute missing data by conditioning on the observed data and model parameters. Usually, if an incomplete data problem permits an EM algorithm implementation, one should be able to set up a Bayesian data augmentation MCMC sampler and vice versa. See Tanner and Wong (1987) for more examples of Bayesian data augmentation.

3.4 Bibliographic remarks

To be added.