Statistics Surveys
ISSN: 1935-7516

Supplement to: A review of Monte Carlo-based versions of the EM algorithm

William Ruth[1],

[1]*Department of Statistics and Actuarial Science, Simon Fraser University, e-mail:* wruth@sfu.ca

In this supplement, we present details for the analysis of our example on estimating gene frequency. See Section 2.2 of the main text for formulation of the model and definition of notation.

1. Observed Data Likelihood, Score and Information

Let π_i be the probability of blood type i. The observed data log-likelihood for our model can be written as follows:

$$\ell(\theta; y) = \log \binom{n}{y} + \sum y_i \log \pi_i(\theta) \tag{1}$$

$$\equiv \sum y_i \log \pi_i \tag{2}$$

$$\equiv 2y_1 \log r + y_2 \log(p^2 + 2pr) + y_3 \log(q^2 + 2qr) + y_4 \log pq \tag{3}$$

where we use $\equiv$ to denote equality up to additive constants which do not depend on θ.

Differentiating ℓ with respect to θ and recalling that $r = 1 - p - q$, so $\partial_p r = \partial_q r = -1$, we get the following expression for the observed data score, S.

$$S(\theta; y) = \begin{pmatrix} \partial_p \ell(\theta; y) \\ \partial_q \ell(\theta; y) \end{pmatrix}, \text{where} \tag{4}$$

$$\partial_p \ell(\theta; y) = -\frac{2y_1}{r} + \frac{2ry_2}{p^2 + 2pr} - \frac{2qy_3}{q^2 + 2qr} + \frac{y_4}{p} \tag{5}$$

$$\partial_q \ell(\theta; Y) = -\frac{2y_1}{r} - \frac{2py_2}{p^2 + 2pr} + \frac{2ry_3}{q^2 + 2qr} + \frac{y_4}{q} \tag{6}$$

Solving the score equation, $S(\theta) = 0$, thus reduces to solving a system of two polynomials in p and q. Since p and q are proportions, we reject any roots outside the unit simplex.

arXiv: 2010.00000

Differentiating ℓ again and multiplying by -1 gives the observed data information matrix, I. To simplify notation, let $p_y = p^2 + 2pr$ and $q_y = q^2 + 2qr$.

$$I(\theta; y) = -\begin{bmatrix} \partial_p^2 \ell(\theta; y) & \partial_{p,q} \ell(\theta; y) \\ \partial_{p,q} \ell(\theta; y) & \partial_q^2 \ell(\theta; y) \end{bmatrix}, \text{where} \tag{7}$$

$$\partial_p^2 \ell(\theta; y) = \frac{2y_1}{r^2} + \frac{2y_2(p_y + 2r^2)}{p_y^2} + \frac{4y_3 q^2}{q_y^2} + \frac{y_4}{p^2} \tag{8}$$

$$\partial_{p,q} \ell(\theta; y) = \frac{2y_1}{r^2} + \frac{2y_2 p^2}{p_y^2} + \frac{2y_3 q^2}{q_y^2} \tag{9}$$

$$\partial_q^2 \ell(\theta; y) = \frac{y_1}{r^2} + \frac{4y_2 p^2}{p_y^2} + \frac{2y_3(q_y + 2r)}{q_y^2} + \frac{y_4}{q^2} \tag{10}$$

The asymptotic standard error of our MLE is I^{-1}, evaluated at the estimate.

2. Complete Data Likelihood, Score and Information

The complete data distribution for our model can be written as follows. Write ρ_i for the probability of genotype i. See Table 2 of the main text for the values of these probabilities.

$$\ell_c(\theta; y, x) = \log \binom{n}{x} + \sum x_i \log \rho_i(\theta) \tag{11}$$

$$\equiv \sum y_i \log \rho_i \tag{12}$$

$$\equiv 2x_1 \log r + x_2 \log pr + 2x_3 \log p + x_4 \log qr + 2x_5 \log q + x_6 \log pq \tag{13}$$

$$= (2x_1 + x_2 + x_4) \log r + (x_2 + 2x_3 + x_6) \log p + (x_4 + 2x_5 + x_6) \log q \tag{14}$$

$$= n_O \log r + n_A \log p + n_B \log q \tag{15}$$

where n_O, n_A and n_B are the number of times allele O, A and B arise respectively in the sampled genotypes. Note that ℓ_c depends on y only through x, so we suppress y from our notation for complete data quantities. The complete data score function is

$$S_c(\theta; x) = \begin{pmatrix} \partial_p \ell_c(\theta; x) \\ \partial_q \ell_c(\theta; x) \end{pmatrix}, \text{where} \tag{16}$$

$$\partial_p \ell_c(\theta; x) = \frac{x_2 + 2x_3 + x_6}{p} - \frac{2x_1 + x_2 + x_4}{r} = \frac{n_A}{p} - \frac{n_O}{r} \tag{17}$$

$$\partial_p \ell_c(\theta; x) = \frac{x_4 + 2x_5 + x_6}{q} - \frac{2x_1 + x_2 + x_4}{r} = \frac{n_B}{q} - \frac{n_O}{r} \tag{18}$$

Notice that the score is linear in x. To make this relationship explicit, we write $S_c(\theta; x) = \mathscr{S}(\theta)x$, where $\mathscr{S}(\theta) \in \mathbb{R}^{2\times 6}$ is a matrix consisting of the coefficients on x in (17) and (18). We will make use of this linearity in Section 5.

Next, we give the information matrix for the complete data.

$$I_c(\theta; x) = -\begin{bmatrix} \partial_p^2 \ell_c(\theta; x) & \partial_{p,q} \ell_c(\theta; x) \\ \partial_{p,q} \ell_c(\theta; x) & \partial_q^2 \ell_c(\theta; x) \end{bmatrix}, \text{where} \tag{19}$$

$$\partial_p^2 \ell_c(\theta; x) = \frac{x_2 + 2x_3 + x_6}{p^2} + \frac{2x_1 + x_2 + x_4}{r^2} = \frac{n_A}{p^2} + \frac{n_O}{r^2} \tag{20}$$

$$\partial_{p,q} \ell_c(\theta; x) = \frac{2x_1 + x_2 + x_4}{r^2} = \frac{n_O}{r^2} \tag{21}$$

$$\partial_q^2 \ell_c(\theta; x) = \frac{x_4 + 2x_5 + x_6}{q^2} + \frac{2x_1 + x_2 + x_4}{r^2} = \frac{n_B}{q^2} + \frac{n_O}{r^2} \tag{22}$$

3. Missing Data Distribution

Many quantities which arise in the EM and MCEM algorithms depend on the missing data distribution (i.e. the conditional distribution of X given $Y = y$). This distribution is best described componentwise in X. First, note that $X_1 = y_1$ and $X_6 = y_4$. Next, we have that $X_2 + X_3 = y_2$ and $X_4 + X_5 = y_3$. Thus, we can write $X_2 | Y = y \sim \text{Bin}(y_2, 2pr/(p^2 + 2pr))$ and $X_4 | Y = y \sim \text{Bin}(y_3, 2qr/(q^2 + 2qr))$. Finally, we recover X_3 and X_5 by subtracting X_2 from y_2 and X_4 from y_3 respectively.

We make frequent use of the first few conditional moments of X, so they are listed here for convenience. Let $\alpha_1 = 2pr/(p^2+2pr)$ be the probability parameter for the binomial distribution of X_2 given Y, and $\alpha_2 = 1 - \alpha_1$. Similarly, let $\beta_1 = 2qr/(q^2 + 2qr)$ correspond to X_4 and $\beta_2 = 1 - \beta_1$.

$$\mathbb{E}(X|Y = y) = \left(y_1, y_2\alpha_1, y_2\alpha_2, y_3\beta_1, y_3\beta_2, y_4\right)^T \tag{23}$$

$$=: \mu_m \tag{24}$$

$$\mathbb{V}(X|Y = y) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & y_2\alpha_1\alpha_2 & -y_2\alpha_1\alpha_2 & 0 & 0 & 0 \\ 0 & -y_2\alpha_1\alpha_2 & y_2\alpha_1\alpha_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & y_3\beta_1\beta_2 & -y_3\beta_1\beta_2 & 0 \\ 0 & 0 & 0 & -y_3\beta_1\beta_2 & y_3\beta_1\beta_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{25}$$

$$=: \Sigma_m \tag{26}$$

$$\mathbb{E}(XX^T|Y = y) = \Sigma_m + \mu_m\mu_m^T \tag{27}$$

Conditional expectations of the number of alleles of each kind will be of partic-

ular interest.

$$\nu_O := \mathbb{E}(n_O|y) \tag{28}$$
$$= 2y_1 + \frac{y_2 pr}{p^2 + 2pr} + \frac{y_3 qr}{q^2 + 2qr} \tag{29}$$
$$= 2y_1 + y_2\left(\frac{\rho_2}{\rho_2 + \rho_3}\right) + y_3\left(\frac{\rho_4}{\rho_4 + \rho_5}\right) \quad \left(= 2y_1 + y_2\left(\frac{\rho_2}{\pi_2}\right) + y_3\left(\frac{\rho_4}{\pi_3}\right)\right) \tag{30}$$

$$\nu_A := \mathbb{E}(n_A|y) \tag{31}$$
$$= \frac{2y_2 pr}{p^2 + 2pr} + \frac{2y_2 p^2}{p^2 + 2pr} + y_4 \tag{32}$$
$$= y_2\left(\frac{\rho_2}{\rho_2 + \rho_3} + \frac{2\rho_3}{\rho_2 + \rho_3}\right) + y_4 \qquad \left(= y_2\left(\frac{\rho_2}{\pi_2} + \frac{2\rho_3}{\pi_2}\right) + y_4\right) \tag{33}$$
$$= y_2\left(1 + \frac{p^2}{p^2 + 2pr}\right) + y_4 \tag{34}$$

$$\nu_B := \mathbb{E}(n_B|y) \tag{35}$$
$$= \frac{2y_3 qr}{q^2 + 2qr} + \frac{2y_3 q^2}{q^2 + 2qr} + y_4 \tag{36}$$
$$= y_3\left(\frac{\rho_4}{\rho_4 + \rho_5} + \frac{2\rho_5}{\rho_4 + \rho_5}\right) + y_4 \qquad \left(= y_3\left(\frac{\rho_4}{\pi_3} + \frac{2\rho_5}{\pi_3}\right) + y_4\right) \tag{37}$$
$$= y_3\left(1 + \frac{q^2}{q^2 + 2qr}\right) + y_4 \tag{38}$$

4. EM Algorithm

In order to apply the EM algorithm, we must construct and optimize the EM objective function. That is, we must compute $Q(\theta|\theta_0) = \mathbb{E}_{\theta_0}\left[\ell_c(\theta; y, X)|Y = y\right]$. The EM objective function can be written as

$$Q(\theta|\theta_0) := \mathbb{E}_{\theta_0}[\ell_c(\theta; X)|Y = y] \tag{39}$$
$$\equiv \nu_O^{(0)} \log r + \nu_A^{(0)} \log p + \nu_B^{(0)} \log q \tag{40}$$

where a superscript zero denotes that the quantity is computed by taking an expectation under θ_0. Differentiating Q with respect to p and q and setting the

result to zero, we get the following system of equations:

$$\frac{\nu_A^{(0)}}{p} = \frac{\nu_O^{(0)}}{r} \tag{41}$$

$$\frac{\nu_B^{(0)}}{q} = \frac{\nu_O^{(0)}}{r} \tag{42}$$

This system of equations can be used to solve for a fixed point of the EM algorithm by evaluating ν_O, ν_A and ν_B at θ instead of θ_0. Note that the fixed point equations which result from this substitution exactly match the observed data score equations given by equations (5) and (6). Indeed, this relationship holds in general under mild conditions (Wu, 1983).

5. Asymptotic Standard Error

Recall that the EM algorithm computes the MLE, which has asymptotic covariance matrix equal to the inverse Fisher information matrix evaluated at the true parameter value. In practice, we estimate this covariance with the inverse of the observed information matrix evaluated at the MLE. Using Proposition 2.3 from the main text, we can calculate the observed information matrix using conditional expectations of quantities derived from the complete data likelihood.

To this end, we need to evaluate the conditional expectations in expression (7) of Proposition 2.3 in the main text. It is convenient for us to write $S_c(\theta) =: \mathscr{S}(\theta)X$ (see Section 2). Then

$$I_c(\hat{\theta}) = \begin{bmatrix} \frac{\nu_A}{p^2} + \frac{\nu_O}{r^2} & \frac{\nu_O}{r^2} \\ \frac{\nu_O}{r^2} & \frac{\nu_B}{q^2} + \frac{\nu_O}{r^2} \end{bmatrix}, \text{and} \tag{43}$$

$$\mathbb{E}_{\hat{\theta}}[S_c(\hat{\theta})S_c(\hat{\theta})^T|Y=y] = \mathscr{S}(\hat{\theta})\mathbb{E}_{\hat{\theta}}\left[XX^T|Y=y\right]\mathscr{S}(\hat{\theta}) \tag{44}$$

$$= \mathscr{S}(\hat{\theta})(\Sigma_m + \mu_M\mu_M^T)\mathscr{S}(\hat{\theta}) \tag{45}$$

$$\tag{46}$$

While it is possible to expand the above expressions, they quickly become too long to easily interpret. We instead leave these as computational formulas and use them as a guide for writing `R` or `Julia` code.

References

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics* **11**.