Integrating the likelihood function of the somatic mutation latent variable model

Louis Dijkstra*

November 10, 2014

The goal is to compute/estimate the integral [1]

$$\int_a^b L(\theta_h, \theta_c \mid Z^h, Z^c) d\theta_c \qquad (\theta_h \text{ is fixed}) \tag{0.1}$$

where

$$L(\theta_h, \theta_c \mid Z^h, Z^c) = \prod_{i=1}^{k} g_i^h \left(Z_i^h \mid \theta_h\right) \times \prod_{j=1}^{l} g_j^c \left(Z_j^c \mid \theta_h, \theta_c; \alpha\right). \tag{0.2}$$

Note that we can rewrite the integral to

$$\underbrace{\prod_{i=1}^{k} g_i^h \left(Z_i^h \mid \theta_h\right)}_{\text{constant}} \times \int_a^b \prod_{j=1}^{l} g_j^c \left(Z_j^c \mid \theta_h, \theta_c; \alpha\right) d\theta_c. \tag{0.3}$$

The conditional probability distribution for observation $g_j^c \left(Z_j^c \mid \theta_h, \theta_c; \alpha\right)$ can be written as:

$$g_j^c \left(Z_j^c \mid \theta_h, \theta_c; \alpha\right) = s_j \theta_c + t_j \tag{0.4}$$

where

$$s_j = (1 - \alpha) \left[p_j^c (Z_j) - a_j^c (Z_j)\right] \tag{0.5}$$

and

$$t_j = \pi_j^c \alpha \left[\theta_h p_j^c (Z_j) + (1 - \theta_h) a_j^c (Z_j)\right] + (1 - \alpha) a_j^c (Z_j) + \left(1 - \pi_j^c\right) u_j^c (Z_j). \tag{0.6}$$

The product $\prod_{j=1}^{l} g_j^c \left(Z_j^c \mid \theta_h, \theta_c; \alpha\right)$ can, therefore, be written as m-order polynomial

$$\prod_{j=1}^{l} (s_j \theta_c + t_j) = \sum_{m=0}^{k} \beta_m \theta_c^m \tag{0.7}$$

with coefficients

$$\beta_m = \sum_{\substack{I \subset \{1,\ldots,l\} \\ |I| = m}} \prod_{i \in I} s_i \prod_{j \notin I} t_j \tag{0.8}$$

*E-mail: `dijkstra@cwi.nl`

[1] The prior $h(\theta_h, \theta_c)$ is initially assumed uniform over the parameter space.

The integral in (0.3) can be written as

$$\int_a^b \sum_{m=0}^{k} \beta_m \theta_c^m d\theta_h = \sum_{m=0}^{k} \frac{\beta_m}{m+1} \left(b^{m+1} - a^{m+1}\right). \tag{0.9}$$

What remains is determining the coefficients $\beta_0, \ldots, \beta_m$ efficiently. Computing them directly from eq. (0.8) is infeasbile, since it requires us to sum over all possible subsets of observations. We propose the following iterative process:

Initialization $\beta_0^{(0)} := 1$

Update steps add every observation $i = 1, 2, \ldots, l$ one by one and 'update' the coefficients with the following rules:

$$\begin{aligned} \beta_0^{(i)} &= \beta_0^{(i-1)} t_i \\ \beta_m^{(i)} &= \beta_m^{(i-1)} t_i + \beta_{m-1}^{i-1} s_i \quad (m = 1, 2 \ldots, i-1) \\ \beta_i^{(i)} &= \beta_{i-1}^{(i-1)} s_i. \end{aligned} \tag{0.10}$$

Termination the coefficients $\beta_1, \ldots, \beta_m$ are

$$\beta_m = \beta_m^{(l)} \text{ for } m = 1, 2, \ldots, l. \tag{0.11}$$

Arbitrary prior distribution In case of an arbitrary prior, the integral

$$\int_a^b L(\theta_h, \theta_c \mid Z^h, Z^c) h(\theta_h, \theta_c) d\theta_c \qquad (\theta_h \text{ is fixed}) \tag{0.12}$$

is equal to

$$\prod_{i=1}^{k} g_i^h \left(Z_i^h \mid \theta_h\right) \times \sum_{m=0}^{l} \beta_m \int_a^b \theta_c^m h(\theta_h, \theta_c) d\theta_c \tag{0.13}$$

where the coefficient β_m can be computed as before.