Estimation of unknown HIV seroconversion time

We assume that a bivariate model, with known parameters,

$$f(\mathbf{y}_i^c, \mathbf{y}_i^r | \mathbf{t}_i^c, \mathbf{t}_i^r) \tag{1}$$

correctly characterizes the evolution of CD4 cell count and HIV-RNA viral load (appropriately transformed; denoted by superscripts c and r,respectively) over time since HIV seroconversion (t) while individuals are ART naive and AIDS free (i.e. during natural history). In seroprevalent individuals, seroconversion date is unknown and times of CD4 cell count and viral load measurements can be expressed relative to their diagnosis date. Denoting these observed times since diagnosis by d, it follows that times of measurements since seroconversion can be expressed as: $t_{ij}^c = d_{ij}^c + w_i$ and $t_{ij}^r = d_{ij}^r + w_i$ with the unknown quantity w_i denoting the time gap between HIV seroconversion and diagnosis dates for the i-th individual.

We can now express the distribution of the biomarkers conditionally on w_i by simply replacing t_{ij}^c and t_{ij}^r with $d_{ij}^c + w_i$ and $d_{ij}^r + w_i$ in (1), respectively. Given the observed measurements $(\mathbf{y}_i^c, \mathbf{y}_i^r)$, we reverse the problem deriving the posterior distribution of the unknown w_i conditionally on $(\mathbf{y}_i^c, \mathbf{y}_i^r)$. This can be easily carried out through Bayes Theorem. Letting $\mathbf{y}_i^\top = (\mathbf{y}_i^c, \mathbf{y}_i^r)$ be the observed measurement of both markers, the posterior distribution of w_i becomes

$$f(w_i|\mathbf{y}_i) = \frac{f(\mathbf{y}_i|w_i)f(w_i)}{\int_0^{u_i} f(\mathbf{y}_i|w_i)f(w_i)dw_i}, \quad 0 < w_i < u_i, \tag{2}$$

where the dependence on the parameters of the measurement model is suppressed for ease of notation, $f(\mathbf{y}_i) = \int_0^{u_i} f(\mathbf{y}_i|w_i)f(w_i)dw_i$ is a normalizing constant and u_i is the upper limit for the possible values of the gap between HIV seroconversion and diagnosis w_i (we assume that HIV seroconversion must occur after the age of 10, after 1/1/1980 and after any documented HIV negative test). We initially assume a uniform prior distribution for w_i over the interval $(0, u_i)$ but for the specific application we update it according to subject-specific behavioral data and AIDS status.

Subject-specific estimates of the unknown gap between HIV seroconversion and diagnosis w_i can be derived through the posterior mean, median or mode whereas the posterior probabilities of HIV acquisition post-migration can be expressed as:

$$\pi_i = \mathrm{P}\left(w_i < m_i\right) = \int_0^{m_i} f(w_i|\mathbf{y}_i)dw_i, \tag{3}$$

where m_i denotes the gap between migration and HIV diagnosis.

The model we used to characterize the natural history evolution of CD4 cell count and HIV-RNA viral load was a bivariate linear mixed model with normally distributed random effects and level-1 residuals. Thus the marginal distribution of the observed measurements over time since HIV infection is the following multivariate normal:

$$\begin{pmatrix} \mathbf{Y}_i^c(\mathbf{t}_i^c) \\ \mathbf{Y}_i^r(\mathbf{t}_i^r) \end{pmatrix} \sim N(\boldsymbol{\mu}_i(\mathbf{t}_i), \mathbf{V}_i(\mathbf{t}_i)), \tag{4}$$

where the mean vector $\boldsymbol{\mu}_i(\mathbf{t}_i)$ and the variance-covariance matrix $\mathbf{V}_i(\mathbf{t}_i)$ are equal to:

$$\boldsymbol{\mu}_i(\mathbf{t}_i) = \begin{pmatrix} \mathbf{X}_i^c(\mathbf{t}_i^c) & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_i^r(\mathbf{t}_i^r) \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}^c \\ \boldsymbol{\beta}^r \end{pmatrix}, \text{ and}$$

$$\mathbf{V}_i(\mathbf{t}_i) = \begin{pmatrix} \mathbf{Z}_i^c(\mathbf{t}_i^c) & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_i^r(\mathbf{t}_i^r) \end{pmatrix} \mathbf{D} \begin{pmatrix} \mathbf{Z}_i^c(\mathbf{t}_i^c) & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_i^r(\mathbf{t}_i^r) \end{pmatrix}^\top + \begin{pmatrix} \sigma_c^2 \mathbf{I}_{n_i^c} & \mathbf{0} \\ \mathbf{0} & \sigma_r^2 \mathbf{I}_{n_i^r} \end{pmatrix} \Bigg),$$

respectively. X and Z denote design matrices for fixed ($\boldsymbol{\beta}$) and random effects, respectively, D is an unstructured variance covariance matrix of random effects of both markers, σ^2 denote level-1 residual variances and I are identidy matrices.

The model was fitted to the CASCADE seroconverters natural history data in which seroconversion dates are well estimated. We used the following variables as covariates: sex, age at infection, region of birth (Africa, Europe, Asia, America), mode of infection and calendar year of infection. Continuous covariates entered the model through splines. CD4 decline was assumed linear after a fourth root transformation and log_{10} transformed viral load was modeled through a fractional polynomial of time which allowed us to capture its non-linear evolution (fast exponential-like decrease within the first months after seroconversion and slow increase thereafter).

Equation (4) was used to express $f(\mathbf{y}_i|w_i)$ in (2) and derive the posterior distribution of the unknown gap between seroconversion and diagnosis w_i. The estimation procedure was built in R using the base functions `optim` and `integrate` along with the additional library `mvtnorm`.