Supplementary Material for "Two-part models with stochastic processes for modelling longitudinal semicontinuous data: computationally efficient inference and modelling the overall marginal mean"

Sean Yiu* and Brian D. M. Tom
*Email-address: Sean.yiu@mrc-bsu.cam.ac.uk
MRC Biostatistics Unit, School of Clinical Medicine,
University of Cambridge, Cambridge, CB2 0SR, UK

Two-part log-normal model

Another approach to modelling longitudinal semicontinuous data is to specify a two-part log-normal model for Y_{ij}. In particular, we assume that the model for the binary component $U_{ij} = I(Y_{ij} > 0)$ is Bernoulli with response probability

$$\mathbb{P}(U_{ij} = 1 | b_{ij}) = \Phi(\boldsymbol{X}_{ij}^{\top}\boldsymbol{\beta} + b_{ij})$$

and that the continuous component follows a log-normal distribution with probability density function

$$f(y_{ij}|Y_{ij} > 0; c_{ij}) = \frac{1}{y_{ij}\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(\log(y_{ij}) - \boldsymbol{Z}_{ij}^{\top}\boldsymbol{\gamma} - c_{ij} + \sigma^2/2)^2}{2\sigma^2}\right\}$$

where $\mathbb{E}[Y_{ij}|Y_{ij} > 0; c_{ij}] = \exp(\boldsymbol{Z}_{ij}^{\top}\boldsymbol{\gamma} + c_{ij})$, $\sigma > 0$ and b_{ij} and c_{ij} are realisations of the stochastic processes described in Sections 3.1 and 3.2 of the main text. Note that in contrast to the main text, the patient-specific mean of the positive values is now modelled using a log-link function and therefore a unit change of covariate values will multiply the patient-specific mean of the positive values by the exponent of the corresponding components in $\boldsymbol{\gamma}$.

Computationally efficient inference for the two-part log-normal model

In this section, we demonstrate how computationally efficient inference for the two-part log-normal model can be obtained. Because the ideas are similar to Section 4 of the main text and consequently the derivations of the results follow in a similar manner, we omit minor details and state only the main results.

For ease of exposition, we describe the likelihood contribution from patient i. The notation and conventions used in this section will be consistent with the main text. For models that contain two stochastic processes, the likelihood contribution from patient i is

$$\begin{aligned} L_i(\boldsymbol{\Theta}) = & \int_{\boldsymbol{b_i}} \int_{\boldsymbol{c_i}} \left[\prod_{j=1}^{m_i} \Phi(\boldsymbol{X}_{ij}^\top \boldsymbol{\beta} + b_{ij})^{u_{ij}} \{1 - \Phi(\boldsymbol{X}_{ij}^\top \boldsymbol{\beta} + b_{ij})\}^{1-u_{ij}} \right] \\ & \times \left[\prod_{j=1}^{m_i} \frac{1}{(y_{ij} + 1 - u_{ij})(\sqrt{2\pi\sigma^2})^{u_{ij}}} \exp\left\{ -\frac{(\log(y_{ij} + 1 - u_{ij}) - u_{ij}(\boldsymbol{Z}_{ij}^\top \boldsymbol{\gamma} + c_{ij} - \sigma^2/2))^2}{2\sigma^2} \right\} \right] \\ & \times \phi^{(2m_i)}(\boldsymbol{b}_i, \boldsymbol{c}_i; \boldsymbol{0}, \boldsymbol{\Sigma}_i)\, d\boldsymbol{b}_i\, d\boldsymbol{c}_i. \end{aligned} \tag{A1}$$

Similarly, for models containing a single stochastic process (i.e. shared process models), the likelihood contribution from patient i is

$$\begin{aligned} L_i(\boldsymbol{\Theta}) = & \int_{\boldsymbol{b_i}} \left[\prod_{j=1}^{m_i} \Phi(\boldsymbol{X}_{ij}^\top \boldsymbol{\beta} + b_{ij})^{u_{ij}} \{1 - \Phi(\boldsymbol{X}_{ij}^\top \boldsymbol{\beta} + b_{ij})\}^{1-u_{ij}} \right] \\ & \times \left[\prod_{j=1}^{m_i} \frac{1}{(y_{ij} + 1 - u_{ij})(\sqrt{2\pi\sigma^2})^{u_{ij}}} \exp\left\{ -\frac{(\log(y_{ij} + 1 - u_{ij}) - u_{ij}(\boldsymbol{Z}_{ij}^\top \boldsymbol{\gamma} + \theta b_{ij} - \sigma^2/2))^2}{2\sigma^2} \right\} \right] \\ & \times \phi^{(m_i)}(\boldsymbol{b}_i; \boldsymbol{0}, \boldsymbol{\Sigma}_{ib})\, d\boldsymbol{b}_i. \end{aligned} \tag{A2}$$

In matrix form, we describe a generic likelihood contribution that encapsulates (A1) and (A2):

$$\begin{aligned} L_i(\boldsymbol{\Theta}) = & \int_{\boldsymbol{l}_i} \Phi^{(m_i)}\left(\boldsymbol{A}_{i1}\boldsymbol{\mu}_{ib} + \boldsymbol{A}_{i2}\boldsymbol{l}_i; \boldsymbol{0}, \boldsymbol{I}_{m_i}\right) \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{\sum_{j=1}^{m_i} u_{ij}} \prod_{j=1}^{m_i} \frac{1}{y_{ij} + 1 - u_{ij}} \\ & \times \exp\left\{ -\frac{(\log(\boldsymbol{y}_i + \boldsymbol{1} - \boldsymbol{u}_i) - \boldsymbol{A}_{i3}(\boldsymbol{\mu}_{ic} - \sigma^2/2) - \boldsymbol{A}_{i4}\boldsymbol{l}_i)^\top (\log(\boldsymbol{y}_i + \boldsymbol{1} - \boldsymbol{u}_i) - \boldsymbol{A}_{i3}(\boldsymbol{\mu}_{ic} - \sigma^2/2) - \boldsymbol{A}_{i4}\boldsymbol{l}_i)}{2\sigma^2} \right\} \\ & \times \phi^{\{\dim(\boldsymbol{l}_i)\}}(\boldsymbol{l}_i; \boldsymbol{0}, \boldsymbol{\Sigma}_{il})\, d\boldsymbol{l}_i \end{aligned} \tag{A3}$$

where $\boldsymbol{l}_i = (\boldsymbol{b}_i, \boldsymbol{c}_i)$, $\boldsymbol{A}_{i2} = (diag(2\boldsymbol{u}_i - \boldsymbol{1}), diag(\boldsymbol{0}))$ and $\boldsymbol{A}_{i4} = (diag(\boldsymbol{0}), diag(\boldsymbol{u}_i))$ for (A1) and $\boldsymbol{l}_i = \boldsymbol{b}_i$, $\boldsymbol{A}_{i2} = diag(2\boldsymbol{u}_i - \boldsymbol{1})$ and $\boldsymbol{A}_{i4} = diag(\theta\boldsymbol{u}_i)$ for (A2). By noting the similarity of (A3) to

(9) in the main text, particularly $g(\boldsymbol{y_i})$ and $\boldsymbol{\mu}_{ic}$ are replaced with $\log(\boldsymbol{y}_i + \mathbf{1} - \boldsymbol{u}_i)$ and $\boldsymbol{\mu}_{ic} - \sigma^2/2$ respectively and that there is an additional multiplicative constant $\prod_{j=1}^{m_i} 1/(y_{ij} + 1 - u_{ij})$, it is clear that (A3) can be re-expressed as

$$L_i(\boldsymbol{\Theta}) = \Phi^{(m_i)}\left(\boldsymbol{A}_{i1}\boldsymbol{\mu}_{ib} + \boldsymbol{A}_{i2}\boldsymbol{h}_i; \mathbf{0}, \boldsymbol{I}_{m_i} + \boldsymbol{A}_{i2}\boldsymbol{H}_i^{-1}\boldsymbol{A}_{i2}^\top\right)\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{\sum_{j=1}^{m_i} u_{ij}} \prod_{j=1}^{m_i} \frac{1}{y_{ij} + 1 - u_{ij}} \frac{1}{|\boldsymbol{\Sigma}_{il}\boldsymbol{H}_i|^{1/2}} \tag{A4}$$

$$\times \exp\left\{-\frac{(\log(\boldsymbol{y}_i + \mathbf{1} - \boldsymbol{u}_i) - \boldsymbol{A}_{i3}(\boldsymbol{\mu}_{ic} - \sigma^2/2))^\top(\log(\boldsymbol{y}_i + \mathbf{1} - \boldsymbol{u}_i) - \boldsymbol{A}_{i3}(\boldsymbol{\mu}_{ic} - \sigma^2/2))}{2\sigma^2} + \frac{\boldsymbol{h}_i^\top \boldsymbol{H}_i \boldsymbol{h}_i}{2}\right\}$$

where

$$\begin{aligned} \boldsymbol{A}_{i1} &= \mathrm{diag}(2\boldsymbol{u}_i - \mathbf{1}) \\ \boldsymbol{A}_{i3} &= \mathrm{diag}(\boldsymbol{u}_i) \\ \boldsymbol{H}_i &= \boldsymbol{A}_{i4}^\top \boldsymbol{A}_{i4}/\sigma^2 + (\boldsymbol{\Sigma}_{il})^{-1} \\ \boldsymbol{h}_i &= \boldsymbol{H}_i^{-1}\boldsymbol{A}_{i4}^\top(\log(\boldsymbol{y}_i + \mathbf{1} - \boldsymbol{u}_i) - \boldsymbol{A}_{i3}(\boldsymbol{\mu}_{ic} - \sigma^2/2))/\sigma^2 \end{aligned} \tag{A5}$$

and $\boldsymbol{\Sigma}_{il} = \boldsymbol{\Sigma}_i$ or $\boldsymbol{\Sigma}_{ib}$ with $\boldsymbol{A}_{i2}$ and $\boldsymbol{A}_{i4}$ defined by the specified model. Because the cumulative distribution function of the multivariate normal distribution is computationally efficient to evaluate, (A4) provides an appealing approach to evaluate the likelihood contributions described by (A1) and (A2).

Modelling the overall marginal mean with the two-part log-normal model

In order to directly model the overall marginal mean, i.e. $\mathbb{E}[Y_{ij}]$, using the two-part log-normal model, we follow the approach of Section 6 by reparameterising the mean of the positive values, i.e. $\mathbb{E}[Y_{ij}|Y_{ij} > 0]$, whilst retaining the same random effects structure that facilitates the computationally efficient implementation procedure. Specifically, following (21) in the main text, let

$$\mathbb{P}(U_{ij} = 1|B_i(t_{ij}) = b_{ij}) = \Phi(\boldsymbol{X}_{ij}^\top\boldsymbol{\beta} + b_{ij})$$
$$\mathbb{E}[Y_{ij}|Y_{ij} > 0, C_i(t_{ij}) = c_{ij}] = \exp(\Delta_{ij} + c_{ij})$$

where Δ_{ij} is a function of covariates (at the jth visit from patient i) and regression coefficients only. The overall marginal mean for the correlated processes and shared process model is given by

$$\mathbb{E}[Y_{ij}] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Phi(\boldsymbol{X}_{ij}^{\top}\boldsymbol{\beta} + b_{ij}) \exp(\Delta_{ij} + c_{ij})\phi^{(2)}(b_{ij}, c_{ij}; \mathbf{0}, \boldsymbol{\Sigma}_{ij})\, db_{ij}\, dc_{ij}$$

and

$$\mathbb{E}[Y_{ij}] = \int_{-\infty}^{\infty} \Phi(\boldsymbol{X}_{ij}^{\top}\boldsymbol{\beta} + b_{ij}) \exp(\Delta_{ij} + \theta b_{ij})\phi^{(1)}(b_{ij}; 0, \sigma_{bij}^2)\, db_{ij}$$

respectively. We begin by computing the overall marginal mean for the shared process model. Firstly, we have

$$\begin{aligned}\mathbb{E}[Y_{ij}] &= \int_{-\infty}^{\infty} \Phi(\boldsymbol{X}_{ij}^{\top}\boldsymbol{\beta} + b_{ij}) \exp(\Delta_{ij} + \theta b_{ij})\phi^{(1)}(b_{ij}; 0, \sigma_{bij}^2)\, db_{ij} \\ &= \exp(\Delta_{ij}) \int_{-\infty}^{\infty}\int_{-\infty}^{\boldsymbol{X}_{ij}^{\top}\boldsymbol{\beta}} \phi(u + b_{ij})\, du \exp(\theta b_{ij})\phi^{(1)}(b_{ij}; 0, \sigma_{bij}^2)\, db_{ij}.\end{aligned}$$

Next, we change the order of integration and complete the square in b_{ij}. This results in

$$\frac{\exp(\Delta_{ij})}{2\pi\sigma_{bij}} \int_{-\infty}^{\boldsymbol{X}_{ij}^{\top}\boldsymbol{\beta}} \int_{-\infty}^{\infty} \exp\left\{ -\frac{1+\sigma_{bij}^2}{2\sigma_{bij}^2}\left(b + \frac{\sigma_{bij}^2(u-\theta)}{1+\sigma_{bij}^2}\right)^2 \right\} \exp\left\{ -\frac{u^2}{2} + \frac{\sigma_{bij}^2(u-\theta)^2}{2(1+\sigma_{bij}^2)} \right\} db_{ij}\, du.$$

After evaluating the above integral with respect to b_{ij} and then completing the square in u, we have

$$\mathbb{E}[Y_{ij}] = \frac{\exp\left\{\Delta_{ij} + \theta^2\sigma_{bij}^2(1+\theta^2\sigma_{bij}^2)/\{2(1+\sigma_{bij}^2)\}\right\}}{\sqrt{2\pi(1+\sigma_{bij}^2)}} \int_{-\infty}^{\boldsymbol{X}_{ij}^{\top}\boldsymbol{\beta}} \exp\left\{ -\frac{(u+\theta\sigma_{bij}^2)^2}{2(1+\sigma_{bij}^2)} \right\} du.$$

The overall marginal mean then results by making the substitution $v = (u + \theta\sigma_{bij}^2)/\sqrt{1+\sigma_{bij}^2}$. Specifically,

$$\begin{aligned}\mathbb{E}[Y_{ij}] &= \exp\left\{ \Delta_{ij} + \frac{\theta^2\sigma_{bij}^2(1+\theta^2\sigma_{bij}^2)}{2(1+\sigma_{bij}^2)} \right\} \int_{-\infty}^{\frac{\boldsymbol{X}_{ij}^{\top}\boldsymbol{\beta}+\theta\sigma_{bij}^2}{\sqrt{1+\sigma_{bij}^2}}} \phi(v)\, dv \\ &= \exp\left\{ \Delta_{ij} + \frac{\theta^2\sigma_{bij}^2(1+\theta^2\sigma_{bij}^2)}{2(1+\sigma_{bij}^2)} \right\} \Phi\left(\frac{\boldsymbol{X}_{ij}^{\top}\boldsymbol{\beta}+\theta\sigma_{bij}^2}{\sqrt{1+\sigma_{bij}^2}} \right)\end{aligned}$$

The overall marginal mean for the correlated processes model can be computed as follows:

$$
\begin{aligned}
\mathbb{E}[Y_{ij}] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Phi(\boldsymbol{X}_{ij}^\top \boldsymbol{\beta} + b_{ij}) \exp(\Delta_{ij} + c_{ij}) \phi^{(2)}(b_{ij}, c_{ij}; \mathbf{0}, \boldsymbol{\Sigma}_{ij})\, db_{ij}\, dc_{ij} \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Phi(\boldsymbol{X}_{ij}^\top \boldsymbol{\beta} + b_{ij}) \exp(\Delta_{ij} + c_{ij}) \phi^{(1)}(b_{ij}; 0, \sigma_{bij}^2) \\
&\times \phi^{(1)}\left(c_{ij}; \frac{\sigma_{cij}\rho_{bcij}b_{ij}}{\sigma_{bij}}, (1-\rho_{bcij}^2)\sigma_{cij}^2\right)\, db_{ij}\, dc_{ij} \\
&= \int_{-\infty}^{\infty} \Phi(\boldsymbol{X}_{ij}^\top \boldsymbol{\beta} + b_{ij}) \phi^{(1)}(b_{ij}; 0, \sigma_{bij}^2) \int_{-\infty}^{\infty} \exp(\Delta_{ij} + c_{ij}) \\
&\quad \times \phi^{(1)}\left(c_{ij}; \frac{\sigma_{cij}\rho_{bcij}b_{ij}}{\sigma_{bij}}, (1-\rho_{bcij}^2)\sigma_{cij}^2\right)\, dc_{ij}\, db_{ij} \\
&= \exp\left\{\frac{(1-\rho_{bcij}^2)\sigma_{cij}^2}{2}\right\} \int_{-\infty}^{\infty} \Phi(\boldsymbol{X}_{ij}^\top \boldsymbol{\beta} + b_{ij}) \exp\left(\Delta_{ij} + \frac{\sigma_{cij}\rho_{bcij}b_{ij}}{\sigma_{bij}}\right) \phi^{(1)}(b_{ij}; 0, \sigma_{bij}^2)\, db_{ij}
\end{aligned}
$$

where the second and third line results from factorizing the bivariate normal density into a conditional and marginal density and the sixth line results from using the moment generating function of the normal distribution. Finally, the overall marginal mean for the correlated processes model can be obtained by noting that the above integral is equivalent to the overall marginal mean of the shared process model with $\theta = \sigma_{cij}\rho_{bcij}/\sigma_{bij}$. Thus we have

$$
\begin{aligned}
\mathbb{E}[Y_{ij}] &= \exp\left\{\frac{(1-\rho_{bcij}^2)\sigma_{cij}^2}{2}\right\} \exp\left\{\Delta_{ij} + \frac{\rho_{bcij}^2\sigma_{cij}^2(1+\rho_{bcij}^2\sigma_{cij}^2)}{2(1+\sigma_{bij}^2)}\right\} \Phi\left(\frac{\boldsymbol{X}_{ij}^\top \boldsymbol{\beta} + \rho_{bcij}\sigma_{bij}\sigma_{cij}}{\sqrt{1+\sigma_{bij}^2}}\right) \\
&= \exp\left\{\Delta_{ij} + \frac{\sigma_{cij}^2}{2}\left(1 - \frac{\rho_{bcij}^2(\sigma_{bij}^2 - \rho_{bcij}^2\sigma_{cij}^2)}{1+\sigma_{bij}^2}\right)\right\} \Phi\left(\frac{\boldsymbol{X}_{ij}^\top \boldsymbol{\beta} + \rho_{bcij}\sigma_{bij}\sigma_{cij}}{\sqrt{1+\sigma_{bij}^2}}\right).
\end{aligned}
$$

It is now implicit that replacing Δ_{ij} with

$$
\boldsymbol{Z}_{ij}^\top \boldsymbol{\alpha} - \frac{\theta^2\sigma_{bij}^2(1+\theta^2\sigma_{bij}^2)}{2(1+\sigma_{bij}^2)} - \log\left\{\Phi\left(\frac{\boldsymbol{X}_{ij}^\top \boldsymbol{\beta} + \theta\sigma_{bij}^2}{\sqrt{1+\sigma_{bij}^2}}\right)\right\}
$$

and

$$
\boldsymbol{Z}_{ij}^\top \boldsymbol{\alpha} - \frac{\sigma_{cij}^2}{2}\left(1 - \frac{\rho_{bcij}^2(\sigma_{bij}^2 - \rho_{bcij}^2\sigma_{cij}^2)}{1+\sigma_{bij}^2}\right) - \log\left\{\Phi\left(\frac{\boldsymbol{X}_{ij}^\top \boldsymbol{\beta} + \rho_{bcij}\sigma_{bij}\sigma_{cij}}{\sqrt{1+\sigma_{bij}^2}}\right)\right\}
$$

in the shared process and correlated processes models respectively will results in $\mathbb{E}[Y_{ij}] = \exp(\boldsymbol{Z}_{ij}^\top \boldsymbol{\alpha})$. Thus the proposed parameterisations allow easily interpretable covariate effects to act on the overall marginal mean, particularly a unit change in covariate values multiplies the overall marginal mean by the exponent of the respective components in $\boldsymbol{\alpha}$, while retaining the proposed computationally efficient implementation procedure.