A joint model and software package for time-to-event and multivariate longitudinal data

G. L. Hickey[1] & P. Philipson[2], A. L. Jorgensen[1], R. Kolamunnage-Dona[1]

[1]*Department of Biostatistics, University of Liverpool, UK*

[2]*Department of Mathematics, Physics and Electrical Engineering, Northumbria University, UK*

February 17, 2020

Vignette summary

A new R package is presented, `joineRML`, which extends the ubiquitous joint model of a single longitudinal measurement outcome and and an event time to the multivariate case of multiple longitudinal outcome types. In this vignette, we outline the *technical* details of the underlying model, estimation algorithm, and ancillary calculations. A separate vignette is available that specifically focuses on application in R.

Contents

1 Model and notation

For each subject $i = 1, \ldots, n$, $y_i = (y_{i1}^\top, \ldots, y_{iK}^\top)$ is the K-variate continuous outcome vector, where each y_{ik} denotes an $(n_{ik} \times 1)$-vector of observed longitudinal measurements for the k-th outcome type: $y_{ik} = (y_{i1k}, \ldots, y_{in_{ik}k})^\top$. Each outcome is measured at observed (possibly pre-specified) times t_{ijk} for $j = 1, \ldots, n_{ik}$, which can differ between subjects and outcomes. Additionally, for each subject there is an event time T_i^*, which is subject to right censoring. Therefore, we observe $T_i = \min(T_i^*, C_i)$, where C_i corresponds to a potential censoring time, and the failure indicator δ_i, which is equal to 1 if the failure is observed $(T_i^* \leq C_i)$ and 0 otherwise. We assume that both censoring and measurement times are non-informative.

The model we describe is an extension of the model proposed by Henderson et al. [1] to the case of multivariate longitudinal data. The model posits an unobserved or latent zero-mean $(K+1)$-variate Gaussian process that is realised independently for each subject, $W_i(t) = \left\{W_{1i}^{(1)}(t), \ldots, W_{1i}^{(K)}(t), W_{2i}(t)\right\}$. This latent process subsequently links the separate sub-models.

The multivariate longitudinal data sub-model, also referred to as the *measurement model* in [1], is given by

$$y_{ik}(t) = \mu_{ik}(t) + W_{1i}^{(k)}(t) + \varepsilon_{ik}(t), \tag{1}$$

where $\mu_{ik}(t)$ is the mean response, and $\varepsilon_{ik}(t)$ is the model error term, which we assume to be independent and identically distributed normal with mean 0 and variance σ_k^2. We assume the mean response is specified as a linear model

$$\mu_{ik}(t) = x_{ik}^\top(t)\beta_k, \tag{2}$$

where $x_{ik}(t)$ is a p_k-vector of (possibly) time-varying covariates with corresponding fixed effect terms β_k. $W_{1i}^{(k)}(t)$ is specified as

$$W_{1i}^{(k)}(t) = z_{ik}^\top(t)b_{ik}, \tag{3}$$

where $z_{ik}(t)$ is an r_k-vector $(r_k \leq p_k)$ of (possibly) time-varying covariates with corresponding subject-and-outcome random effect terms b_{ik}, which follow a zero-mean multivariate normal distribution with $(r_k \times r_k)$-variance-covariance matrix D_{kk}. To account for dependence between the different longitudinal outcome types, we let $\text{Cov}(b_{ik}, b_{il}) = D_{kl}$ for $k \neq l$. Furthermore, we assume $\varepsilon_{ik}(t)$ and b_{ik} are uncorrelated, and that the censoring times are independent of the random effects; both are standard modelling assumptions. These distributional assumptions together with the model given by (1)–(3) is equivalent to the multivariate extension of the Laird and Ware [2] mixed linear effects model. Henderson et al. [1] note that more flexible specifications of $W_{1i}^{(k)}(t)$ can be used, including, for example, stationary Gaussian processes. We do not consider these cases here.

The time-to-event sub-model, also referred to as the *intensity process model* in [1], is given by the hazard function

$$\lambda_i(t) = \lambda_0(t) \exp\left\{v_i^\top(t)\gamma_v + W_{2i}(t)\right\}, \tag{4}$$

where $\lambda_0(\cdot)$ is an unspecified baseline hazard function, and $v_i(t)$ is a q-vector of (possibly) time-varying covariates with corresponding fixed effect terms γ_v. Conditional on $W_i(t)$ and the observed covariate data, the separate data generating processes are *conditionally independent*. To establish a latent association, we specify $W_{2i}(t)$ as a linear combination of $\left\{W_{1i}^{(1)}(t), \ldots, W_{1i}^{(K)}(t)\right\}$

$$W_{2i}(t) = \sum_{k=1}^{K} \gamma_{yk} W_{1i}^{(k)}(t),$$

where $\gamma_y = (\gamma_{y1}, \dots, \gamma_{yK})$ are the joint model association parameters. To emphasise the dependence of $W_{2i}(t)$ on the random effects, we will explicitly write it as $W_{2i}(t, b_i)$ from here onwards. As noted above for $W_{1i}^{(k)}(t)$, $W_{2i}(t)$ can also be flexibly extended, for example to include subject-specific frailty effects [1].

2 Estimation

2.1 Likelihood

For each subject, collect covariate data from different outcome types and measurement times together, denoted as

$$X_i = \begin{pmatrix} X_{i1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & X_{iK} \end{pmatrix}, \qquad Z_i = \begin{pmatrix} Z_{i1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & Z_{iK} \end{pmatrix},$$

where $X_{ik} = \left(X_{i1k}^\top, \dots, X_{in_{ik}k}^\top\right)$ is an $(n_{ik} \times p_k)$-design matrix, with the j-th row corresponding to the p_k-vector of covariates measured at time t_{ijk}. The notation similarly follows for Z_{ik}. Also, collect variance-covariance terms for the random effects and error terms together, denoted as

$$D = \begin{pmatrix} D_{11} & \cdots & D_{1K} \\ \vdots & \ddots & \vdots \\ D_{1K}^\top & \cdots & D_{KK} \end{pmatrix}, \qquad \Sigma_i = \begin{pmatrix} \sigma_1^2 I_{n_{i1}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_K^2 I_{n_{iK}} \end{pmatrix},$$

where I_n denotes an $n \times n$ identity matrix. We further define $\beta = (\beta_1^\top, \dots, \beta_K^\top)$ and $b_i = (b_{i1}^\top, \dots, b_{iK}^\top)$. We can then rewrite the longitudinal data sub-model as

$$\begin{aligned} y_i \mid b_i, \beta, \Sigma_i &\sim N(X_i\beta + Z_i b_i, \Sigma_i), \\ \text{with } b_i \mid D &\sim N(0, D). \end{aligned}$$

For the estimation, we will assume that the covariates in the time-to-event model are time-independent, i.e. $v_i(t) \equiv v_i$. Extensions of the estimation procedure for time-varying covariates are outlined in Rizopoulos [3]. The *observed* data likelihood is given by

$$\prod_{i=1}^{n} \left(\int_{-\infty}^{\infty} f(y_i \mid b_i, \theta) f(T_i, \delta_i \mid b_i, \theta) f(b_i \mid \theta) db_i \right) \tag{5}$$

where $\theta = (\beta^\top, \text{vech}(D), \sigma_1^2, \dots, \sigma_K^2, \lambda_0(t), \gamma_v^\top, \gamma_y^\top)$ is the collection of unknown parameters that we want to estimate, and

$$\begin{aligned} f(y_i \mid b_i, \theta) &= \left(\prod_{k=1}^{K} (2\pi)^{-\frac{n_{ik}}{2}} \right) |\Sigma_i|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (y_i - X_i\beta - Z_i b_i)^\top \Sigma_i^{-1} (y_i - X_i\beta - Z_i b_i) \right\}, \\ f(T_i, \delta_i \mid b_i; \theta) &= \left[\lambda_0(T_i) \exp\left\{ v_i^\top \gamma_v + W_{2i}(T_i, b_i) \right\} \right]^{\delta_i} \exp\left\{ - \int_0^{T_i} \lambda_0(u) \exp\left\{ v_i^\top \gamma_v + W_{2i}(u, b_i) \right\} du \right\}, \\ f(b_i \mid \theta) &= (2\pi)^{-\frac{r}{2}} |D|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} b_i^\top D^{-1} b_i \right\}, \end{aligned}$$

where $r = \sum_{k=1}^{K} r_k$ is the total dimensionality of the random effects variance-covariance matrix.

2.2 EM algorithm

We determine maximum likelihood estimates of the parameters θ using the expectation-maximisation (EM) algorithm [4], by treating the random effects b_i as missing data. This algorithm has been described by Wulfsohn and Tsiatis [5] and Ratcliffe et al. [6] in the context of univariate data joint modelling, and by Lin et al. [7] for multivariate data joint modelling. Starting from an initial estimate of the parameters, the procedure involves iterating between the following two steps until convergence is achieved.

1. *E-step.* At the m-th iteration, we compute the expected log-likelihood of the *complete* data conditional on the *observed* data and the current estimate of the parameters.

$$\begin{aligned} Q(\theta \,|\, \hat{\theta}^{(m)}) &= \sum_{i=1}^{n} \mathbb{E}\Big\{ \log f(y_i, T_i, \delta_i, b_i \,|\, \theta) \Big\}, \\ &= \sum_{i=1}^{n} \int_{-\infty}^{\infty} \Big\{ \log f(y_i, T_i, \delta_i, b_i \,|\, \theta) \Big\} f(b_i \,|\, T_i, \delta_i, y_i; \hat{\theta}^{(m)}) db_i \end{aligned}$$

   Here, the complete-data likelihood contribution for subject i is given by the integrand of (5).

2. *M-step.* We maximise $Q(\theta \,|\, \hat{\theta}^{(m)})$ with respect to θ. namely,

$$\hat{\theta}^{(m+1)} = \arg \max_{\theta} Q(\theta \,|\, \hat{\theta}^{(m)})$$

The updates require expectations about the random effects be calculated of the form $\mathbb{E}\Big[h(b_i) \,|\, T_i, \delta_i, y_i; \hat{\theta}^{(m)}\Big]$, which, in the interests of brevity, we denote here onwards as $\mathbb{E}\left[h(b_i)\right]$ in the update estimates. This expectation is conditional on the observed data (T_i, δ_i, y_i) for each subject, the covariates (including measurement times) (X_i, Z_i, v_i), which are implicitly dependent, and an estimate of the model parameters θ.

2.2.1 M-step details

The M-step estimators naturally follow from Wulfsohn and Tsiatis [5] and Lin et al. [7]. The baseline hazard is estimated in closed-form using the Breslow estimator, with jump size:

$$\hat{\lambda}_0(t) = \frac{\sum_{i=1}^{n} \delta_i I(T_i = t)}{\sum_{i=1}^{n} \mathbb{E}\left[\exp\left\{v_i^\top \gamma_v + W_{2i}(t, b_i)\right\}\right] I(T_i \geq t)}, \tag{6}$$

which is only evaluated a distinct observed event times, t_j $(j = 1, \ldots, J)$, where $I(\mathcal{A})$ denotes an indicator function that takes the value 1 if event $\mathcal{A}$ occurs, and zero otherwise. Updates for β, D, and σ_k^2 (for

$k = 1, \dots, K$) are also given in closed-form as:

$$\begin{aligned}
\hat{\beta} &= \left(\sum_{i=1}^{n} X_i^\top \Sigma_i^{-1} X_i \right)^{-1} \left(\sum_{i=1}^{n} X_i^\top \Sigma_i^{-1} (y_i - Z_i \mathbb{E}[b_i]) \right), \\
&= \left(\sum_{i=1}^{n} X_i^\top X_i \right)^{-1} \left(\sum_{i=1}^{n} X_i^\top (y_i - Z_i \mathbb{E}[b_i]) \right), \\
\hat{\sigma}_k^2 &= \frac{1}{\sum_{i=1}^{n} n_{ik}} \sum_{i=1}^{n} \mathbb{E} \left\{ (y_{ik} - X_{ik}\beta_k - Z_{ik}b_{ik})^\top (y_{ik} - X_{ik}\beta_k - Z_{ik}b_{ik}) \right\}, \\
&= \frac{1}{\sum_{i=1}^{n} n_{ik}} \sum_{i=1}^{n} \left\{ (y_{ik} - X_{ik}\beta_k)^\top (y_{ik} - X_{ik}\beta_k - 2Z_{ik}\mathbb{E}[b_{ik}]) + \text{trace}\left(Z_{ik}^\top Z_{ik} \mathbb{E}[b_{ik}b_{ik}^\top] \right) \right\}, \text{ and} \\
\hat{D} &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[b_i b_i^\top \right].
\end{aligned}$$

The simplification in the estimate for $\hat{\beta}$ derives from X_i and Σ_i being block-diagonal, with the latter being $\text{diag}(\sigma_1^2 I_{n_{i1}}, \dots, \sigma_K^2 I_{n_{iK}})$. The updates for γ_v and γ_y are not available in closed-form, so are updated jointly using a one-step multidimensional Newton-Raphson algorithm as

$$\hat{\gamma}^{(m+1)} = \hat{\gamma}^{(m)} + I\left(\hat{\gamma}^{(m)} \right)^{-1} S\left(\hat{\gamma}^{(m)} \right),$$

where $\hat{\gamma}^{(m)}$ denotes the value of $\gamma = (\gamma_v^\top, \gamma_y^\top)$ at the current iteration, $S\left(\hat{\gamma}^{(m)} \right)$ is the corresponding score function, and $I\left(\hat{\gamma}^{(m)} \right)$ is the observed information matrix, which is equal to the derivative of the negative score function. Further details of this update are given in Appendix A.

2.2.2 E-step details

We calculate the conditional expectation of a function of the random effects as

$$\mathbb{E}\left[h(b_i) \,|\, T_i, \delta_i, y_i; \hat{\theta} \right] = \frac{\int_{-\infty}^{\infty} h(b_i) f(b_i \,|\, y_i; \hat{\theta}) f(T_i, \delta_i \,|\, b_i; \hat{\theta}) db_i}{\int_{-\infty}^{\infty} f(b_i \,|\, y_i; \hat{\theta}) f(T_i, \delta_i \,|\, b_i; \hat{\theta}) db_i}, \tag{7}$$

where $f(T_i, \delta_i \,|\, b_i; \hat{\theta})$ is given by (14), and $f(b_i \,|\, y_i; \hat{\theta})$ is calculated from multivariate normal distribution theory as

$$b_i \,|\, y_i, \theta \sim N\left(A_i \left\{ Z_i^\top \Sigma_i^{-1} (y_i - X_i\beta) \right\}, A_i \right), \tag{8}$$

where $A_i = \left(Z_i^\top \Sigma_i^{-1} Z_i + D^{-1} \right)^{-1}$. The derivation[1] is given in Wulfsohn and Tsiatis [5] and Lin et al. [7]. Gaussian quadrature, notably Gauss-Hermite quadrature, is a standard approach for evaluating the integrals in this estimation approach. However, for multivariate longitudinal data, the additional random effects required is commensurate with an exponential growth in the number of quadrature points at which the integrand must be evaluated. Therefore, we use Monte Carlo sampling methods to evaluate the integrals, which was also used by Lin et al. [7]. For N Monte Carlo sample draws, $\{b_i^{(1)}, b_i^{(2)}, \dots, b_i^{(N)}\}$, (7) is approximated

[1] The formulae given in Wulfsohn and Tsiatis [5] and Lin et al. [7] are equivalent, which can be seen by applying the Woodbury matrix identity.

by the ratio of the sample means for $h(b_i)f(T_i, \delta_i \,|\, b_i; \hat{\theta})$ and $f(T_i, \delta_i \,|\, b_i; \hat{\theta})$ evaluated at these values. Namely,

$$\mathbb{E}\left[h(b_i) \,|\, T_i, \delta_i, y_i; \hat{\theta}\right] \approx \frac{\frac{1}{N}\sum_{d=1}^{N} h\left(b_i^{(d)}\right) f\left(T_i, \delta_i \,|\, b_i^{(d)}; \hat{\theta}\right)}{\frac{1}{N}\sum_{d=1}^{N} f\left(T_i, \delta_i \,|\, b_i^{(d)}; \hat{\theta}\right)}. \tag{9}$$

As proposed by Henderson et al. [1], we will use antithetic simulation for variance reduction in the Monte Carlo integration. Instead of directly sampling from (8), we sample $\Omega \sim N(0, I_r)$ and obtain the *pairs*

$$A_i\left\{Z_i^\top \Sigma_i^{-1}(y_i - X_i\beta)\right\} \pm C_i\Omega,$$

where C_i is the Cholesky decomposition of A_i such that $C_iC_i^\top = A_i$. Therefore we only need to draw $N/2$ samples using this approach, and by virtue of the negative correlation between the pairs, it leads to a smaller variance in the sample means taken in (9) than would be obtained from N independent simulations. The choice of N is described below.

2.2.3 Initial values

The EM algorithm requires that initial parameters are specified, namely $\hat{\theta}^{(0)}$. By choosing values close to the maximizer, the number of iterations required to reach convergence will be reduced.

For the time-to-event sub-model, a quasi-two-stage model is fitted when the measurement times are balanced., i.e. when $t_{ijk} = t_{ij}\ \forall k$ That is, we fit *separate* LMMs for each longitudinal outcome as per (1), ignoring the correlation between different outcomes. This is straightforward to implement using standard software. From the fitted models, the best linear unbiased predictions of the separate model random effects are used to estimate each $W_{1i}^{(k)}(t)$ function. These estimates are then included as time-varying covariates in a Cox regression model, alongside any other fixed effect covariates, which can be straightforwardly fitted using standard software. In the situation that the data are not balanced, i.e. when $t_{ijk} \neq t_{ij}\ \forall k$, then we fit a standard Cox proportional hazards regression model for the baseline covariates, and set $\gamma_{yk} = 0\ \forall k$.

For the longitudinal data sub-model, when $K > 1$ we first find the maximum likelihood estimate of $\{\beta, \text{vech}(D), \sigma_1^2, \dots, \sigma_K^2\}$ by running an EM algorithm for the multivariate linear mixed model. The E- and M-step updates are available in closed form, and the initial parameters for this EM algorithm are available from the separate LMM fits. As these are estimated using an EM rather than MCEM algorithm, we can specify a stricter convergence criterion on the estimates.

2.2.4 Convergence and stopping rules

Two standard stopping rules for the deterministic EM algorithm used to declare convergence are the relative and absolute differences, defined as

$$\Delta_{\text{rel}}^{(m+1)} = \max\left\{\frac{|\hat{\theta}^{(m+1)} - \hat{\theta}^{(m)}|}{|\hat{\theta}^{(m)}| + \epsilon_1}\right\} < \epsilon_0, \text{ and} \tag{10}$$

$$\Delta_{\text{abs}}^{(m+1)} = \max\left\{|\hat{\theta}^{(m+1)} - \hat{\theta}^{(m)}|\right\} < \epsilon_2 \tag{11}$$

respectively, for some appropriate choice of ϵ_0, ϵ_1, and ϵ_2. For reference, the `JM` package implements the relative difference stopping rule (in combination with another rule based on relative change in the likelihood), whereas the `joineR` package implements the absolute difference stopping rule.

The choice of N and the monitoring of convergence are conflated when applying a Monte Carlo EM algorithm (MCEM), and a dynamic approach is required. As noted by Wei and Tanner [8], it is computationally inefficient to use a large N in the early phase of the MCEM algorithm when the parameter estimates are likely to be far from the maximizer. On the flip side, as the parameter estimates approach the maximizer, the above stopping rules will fail as the changes in parameter estimates will be swamped by Monte Carlo error. Therefore, it has been recommended that one increase N as the estimate moves towards the maximizer. Although this might be done subjectively [9] or by pre-specified rules [10], an automated approach is preferable. Booth and Hobert [11] proposed an update rule based on a confidence ellipsoid for the maximizer at the $(m+1)$-th iteration, calculated using an approximate sandwich estimator for the maximizer, which accounts for the Monte Carlo error at each iteration. This approach requires additional variance estimation at each iteration, therefore we opt for a simpler approach described by Ripatti et al. [12]. We calculate a coefficient of variation at the $(m+1)$-th iteration as

$$\mathrm{cv}(\Delta_{\mathrm{rel}}^{(m+1)}) = \frac{\mathrm{sd}(\Delta_{\mathrm{rel}}^{(m-1)}, \Delta_{\mathrm{rel}}^{(m)}, \Delta_{\mathrm{rel}}^{(m+1)})}{\mathrm{mean}(\Delta_{\mathrm{rel}}^{(m-1)}, \Delta_{\mathrm{rel}}^{(m)}, \Delta_{\mathrm{rel}}^{(m+1)})},$$

where $\Delta_{\mathrm{rel}}^{(m+1)}$ is given by (10), and $\mathrm{sd}(\cdot)$ and $\mathrm{mean}(\cdot)$ are the sample standard deviation and mean functions, respectively. If $\mathrm{cv}(\Delta_{\mathrm{rel}}^{(m+1)}) > \mathrm{cv}(\Delta_{\mathrm{rel}}^{(m)})$, then $N := N + \lfloor N/\delta \rfloor$, for some small positive integer δ. Typically, we run the MCEM algorithm with a small N (e.g. a default of $50K$ iterations) before implementing this update rule in order to get into the approximately correct parameter region. Appropriate values for other parameters will be application specific, however we have found $\delta = 3$, $N = 100$, $\epsilon_1 = 0.001$, and $\epsilon_0 = \epsilon_2 = 0.005$ to deliver reasonably accurate estimates.

As the monotonicity property is lost due to the Monte Carlo integrations in MCEM, convergence might be prematurely declared due to stochasticity if the ϵ values are too large. To reduce the chance of this occurring, we require that the stopping rule is satisfied for 3 consecutive iterations [11, 12]. However, in any case, trace plots should be inspected to confirm convergence is appropriate.

2.2.5 Likelihood evaluation

The observed data likelihood is calculated following the observation by Henderson et al. [1] that it can be rewritten as

$$\prod_{i=1}^{n} f(y_i \mid \hat{\theta}) \left(\int_{-\infty}^{\infty} f(T_i, \delta_i \mid b_i, \hat{\theta}) f(b_i \mid y_i, \hat{\theta}) db_i \right),$$

where marginal distribution $f(y_i \mid \theta)$ is a multivariate normal density with mean $X_i \beta$ and variance-covariance matrix $\Sigma_i + Z_i D Z_i^\top$, $f(b_i \mid y_i, \theta)$ is given by (8), and $\hat{\theta}$ is the maximum likelihood estimate determined from the EM algorithm. Once calculated, the Akaike information criterion (AIC) and Bayesian information criterion (BIC) are straightforwardly calculated, which are useful for model selection.

2.3 Standard error estimation

We consider two standard error (SE) estimators:

1. **Approximate SEs.** These are estimated by $I_e^{-1/2}(\hat{\theta})$, where $I_e(\theta)$ is the observed empirical information matrix [13], given by

$$I_e(\theta) = \sum_{i=1}^{n} s_i(\theta) s_i^\top(\theta) - \frac{1}{n} S(\theta) S^\top(\theta), \tag{12}$$

$s_i(\theta)$ is the conditional expectation of the complete-data score function for subject i, and $S(\theta)$ is the score defined by $S(\theta) = \sum_{i=1}^n s_i(\theta)$. At the maximizer, $S(\hat{\theta}) = 0$, meaning that the right hand-side of (12) is zero. Due to the Monte Carlo error in the MCEM algorithm, this will not be exactly zero, and therefore we include it in the calculations.

Owing to the estimation approach described in Appendix A, we calculate approximate SEs for $\theta_{-\lambda_0} = (\beta^\top, \text{vech}(D), \sigma_1^2, \dots, \sigma_K^2, \gamma^\top)$ using the profile score vectors

$$s_i\left(\theta_{-\lambda_0}, \hat{\lambda}_0(t \,|\, \theta_{-\lambda_0})\right),$$

where $\lambda_0(t)$ is substituted by $\hat{\lambda}_0(t)$, which is a function of γ (and implicitly, the other parameters via the conditional expectation in the E-step, (7)) given by (6). Therefore, we do not calculate approximate SEs for the baseline hazard; however, this is generally not of inferential interest, hence the application of the Cox model formulation in the first place.

2. **Bootstrap estimated SEs.** These are estimated by sampling n subjects with replacement and re-labelling the subjects with indices $i' = 1, \dots, n$. We then re-fit the model to the bootstrap-sampled dataset. It is important to note that we re-sample patients, not individual data points. This is repeated B-times. For each iteration, we extract the model parameter estimates for $(\beta^\top, \text{vech}(D), \sigma_1^2, \dots, \sigma_K^2, \gamma_v^\top, \gamma_y^\top)$. Note that as the the event times might be tied, the dimension of $\lambda_0(t)$ evaluated at the unique failure times will vary for each iteration; therefore, for simplicity, we do not consider standard error estimates of the baseline hazard. When B is sufficiently large, the SEs can be estimated from the estimated coefficients. Alternatively, $100(1-\alpha)$%-confidence intervals can be estimated from the the $100\alpha/2$-th and $100(1-\alpha/2)$-th percentiles.

From a theoretical perspective, the preferred SE estimates are those using the bootstrap method, owing to the fact that $\hat{\lambda}_0(t)$ will generally be a high-dimensional vector, which might lead to numerical difficulties in the inversion of the observed information matrix [3], and also because the profile likelihood estimates based on the usual observed information matrix approach are known to be underestimated [14]. The reason for this is that the profile estimates are implicit, since the posterior expectations, given by (7), depend on the parameters being estimated, including $\lambda_0(t)$ [3, page 67]. On the flip side, the bootstrap estimates are computationally expensive. Nonetheless, at the model development stage, it is often of interest to gauge the strength of association of model covariates. For this reason, we propose use of the approximate SE estimator, which were also calculated by Lin et al. [7]. If computationally feasible, however, we recommend that bootstrap SEs are also estimated and contrasted to the approximate ones. We also note that recently it has been suggested that bootstrap estimators *overestimate* the SEs; e.g. Xu et al. [15, p. 740] and Hsieh et al. [14, p. 1041].

As noted earlier, SE estimates other than those deriving from bootstrap estimation are expected to be underestimated. On the other hand, the additional Monte Carlo error will likely lead to larger standard error estimates, whether using either the approximate and bootstrap approach. The consequences of these competing factors operating to both inflate and deflate the standard error estimators is not fully understood, and therefore they must be interpreted with a degree of caution.

3 Simulation

To assess properties of the model estimation algorithm, for example bias and coverage, it is necessary to simulate data from joint models of multivariate longitudinal data and time-to-event data. We consider simulation from two models with either

1. $W_{1i}^{(k)}(t) = b_{ik,0}$ (random-intercepts model), or

2. $W_{1i}^{(k)}(t) = b_{ik,0} + b_{1k,1}t$ (random-intercepts and random-slopes model).

Conditional on model parameters, simulation of complete longitudinal data for subjects is trivial by sampling from multivariate normal distributions. In practice, it is preferable to simulate data for fixed time points $0, 1, \dots, T_{\max}$.

The event times are simulated from an exponential distribution in the case of the random-intercepts model, or Gompertz distribution in the case of a random-intercepts and random-slopes model. In the case of (1), this is equivalent to simulation of event times conditional on known baseline covariates. In the case of (2), this is equivalent to simulation of event times conditional on a linear time-varying covariate. The approaches to simulating event times under each of these situations is reported in Bender et al. [16] and Austin [17], respectively. To illustrate (2), we first note that we can re-write (4) as

$$\begin{aligned} & \lambda_0(t) \exp\left\{ \left(v_i^\top(t)\gamma_v + \sum_{k=1}^{K} \gamma_{yk} b_{ik,0} \right) + \left(\sum_{k=1}^{K} \gamma_{yk} b_{ik,1} \right) t \right\} \\ \equiv \quad & \lambda_0(t) \exp\left\{ s_i + w_i t \right\}. \end{aligned}$$

Hence, using the formula (6) from Austin [17], we can simulate an event time as

$$T_i = \frac{1}{w_i + \theta_0} \log\left(1 - \frac{(w_i + \theta_0)\log(u_i))}{\theta_1 \exp(s_i)} \right),$$

where $u_i \sim U(0, 1)$, and $\theta_0 > 0$ and $-\infty < \theta_1 < \infty$ are the scale and shape parameters of the Gompertz distribution. See Austin [17] for a derivation of this formula. For all simulations, we also consider one continuous covariate and one binary covariate, which can be included in both sub-models.

In practice, we will also observe independent right-censoring. For this, we simulate a censoring time C_i from an exponential distribution with scale $\lambda > 0$, and return the observed data of follow-up time $\min(T_i, C_i)$ and event indicator $I(T_i \leq C_i)$. Additionally, as studies are generally terminated after a fixed follow-up period, we will have a truncation time.

References

[1] R. Henderson, P. J. Diggle, and A. Dobson. "Joint modelling of longitudinal measurements and event time data". In: *Biostatistics* 1.4 (2000), pp. 465–480.

[2] N. M. Laird and J. H. Ware. "Random-effects models for longitudinal data". In: *Biometrics* 38.4 (1982), pp. 963–74.

[3] D. Rizopoulos. *Joint Models for Longitudinal and Time-to-Event Data, with Applications in R*. Boca Raton, FL: Chapman & Hall/CRC, 2012.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 39.1 (1977), pp. 1–38.

[5] M. S. Wulfsohn and A. A. Tsiatis. "A joint model for survival and longitudinal data measured with error". In: *Biometrics* 53.1 (1997), pp. 330–339.

[6] S. J. Ratcliffe, W. Guo, and T. R. Ten Have. "Joint modeling of longitudinal and survival data via a common frailty". In: *Biometrics* 60.4 (2004), pp. 892–899.

[7] H. Lin, C. E. McCulloch, and S. T. Mayne. "Maximum likelihood estimation in the joint analysis of time-to-event and multiple longitudinal variables". In: *Statistics in Medicine* 21 (2002), pp. 2369–2382.

[8] G. C. Wei and M. A. Tanner. "A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms". In: *Journal of the American Statistical Association* 85.411 (1990), pp. 699–704.

[9] N. J. Law, J. M. Taylor, and H Sandler. "The joint modeling of a longitudinal disease progression marker and the failure time process in the presence of cure". In: *Biostatistics* 3.4 (2002), pp. 547–563.

[10] C. E. McCulloch. "Maximum likelihood algorithms for generalized linear mixed models". In: 92.437 (1997), pp. 162–170.

[11] J. G. Booth and J. P. Hobert. "Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.1 (1999), pp. 265–285.

[12] S. Ripatti, K. Larsen, and J. Palmgren. "Maximum likelihood inference for multivariate frailty models using an automated Monte Carlo EM algorithm". In: *Lifetime Data Analysis* 8.2002 (2002), pp. 349–360.

[13] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions.* Second. Wiley-Interscience, 2008.

[14] F. Hsieh, Y. K. Tseng, and J. L. Wang. "Joint modeling of survival and longitudinal data: Likelihood approach revisited". In: *Biometrics* 62.4 (2006), pp. 1037–1043.

[15] C. Xu, P. D. Baines, and J. L. Wang. "Standard error estimation using the EM algorithm for the joint modeling of survival and longitudinal data". In: *Biostatistics* 15.4 (2014), pp. 731–44.

[16] R. Bender, T. Augustin, and M. Blettner. "Generating survival times to simulate Cox proportional hazards models". In: *Statistics in Medicine* 24 (2005), pp. 1713–1723.

[17] P. C. Austin. "Generating survival times to simulate Cox proportional hazards models with time-varying covariates". In: *Statistics in Medicine* 31.29 (2012), pp. 3946–3958.

[18] H. Lin. "A finite mixture model for joint longitudinal biomarker and survival outcome responses". Ph.D. Cornell University, 2000.

A Appendix: Score equations

From (5), the expected complete-data log-likelihood is given by

$$\sum_{i=1}^{n} \int_{-\infty}^{\infty} \Big\{ \log f(y_i, T_i, \delta_i, b_i \,|\, \theta) \Big\} f(b_i \,|\, T_i, \delta_i, y_i, \hat{\theta}^{(m)}) db_i$$

where the expectation is taken over the conditional random effects distribution $f(b_i \,|\, T_i, \delta_i, y_i, \hat{\theta}^{(m)})$.

We can decompose the complete-data log-likelihood for subject i into

$$\log f(y_i, T_i, \delta_i, b_i \,|\, \theta) = \log f(y_i \,|\, b_i, \theta) + \log f(T_i, \delta_i \,|\, b_i, \theta) + \log f(b_i \,|\, \theta),$$

where

$$
\begin{aligned}
\log f(y_i \,|\, b_i, \theta) &= -\frac{1}{2}\left\{\left(\sum_{k=1}^{K} n_{ik}\right)\log(2\pi) + \log|\Sigma_i| + (y_i - X_i\beta - Z_i b_i)^\top \Sigma_i^{-1}(y_i - X_i\beta - Z_i b_i)\right\} && (13)\\
\log f(T_i, \delta_i \,|\, b_i, \theta) &= \delta_i \log\lambda_0(T_i) + \delta_i\left[v_i^\top\gamma_v + W_{2i}(T_i, b_i)\right] - \int_0^{T_i} \lambda_0(u)\exp\left\{v_i^\top\gamma_v + W_{2i}(u, b_i)\right\} du && (14)\\
\log f(b_i \,|\, \theta) &= -\frac{1}{2}\left\{r\log(2\pi) + \log|D| + b_i^\top D^{-1} b_i\right\}. && (15)
\end{aligned}
$$

The update equations are then estimated according to the score equations, $\partial Q(\theta \,|\, \hat{\theta}^{(m)})/\partial\theta$. The score equations are effectively given in Lin et al. [7], although there the random effects were hierarchically centred about the corresponding fixed effect terms as part of a current values parametrisation, as well as being embedded in a frailty Cox model, which has consequences on the score equations here. The score equations for $\lambda_0(t)$, β, and σ_k^2 are

$$
\begin{aligned}
S(\lambda_0(t)) &= \sum_{i=1}^{n}\left\{\frac{\delta_i I(T_i = t)}{\lambda_0(t)} - \mathbb{E}\left[\exp\{v_i^\top\gamma_v + W_{2i}(t, b_i)\}\right] I(T_i \geq t)\right\},\\
S(\beta) &= \sum_{i=1}^{n}\left\{X_i^\top \Sigma_i^{-1}(y_i - X_i\beta - Z_i\mathbb{E}[b_i])\right\},\\
S(\sigma_k^2) &= -\frac{1}{2\sigma_k^2}\sum_{i=1}^{n}\left\{n_{ik} - \frac{1}{\sigma_k^2}\mathbb{E}\left[(y_{ik} - X_{ik}\beta_k - Z_{ik}b_{ik})^\top(y_{ik} - X_{ik}\beta_k - Z_{ik}b_{ik})\right]\right\}\\
&= -\frac{1}{2\sigma_k^2}\sum_{i=1}^{n}\Bigg\{n_{ik} - \frac{1}{\sigma_k^2}\Big[(y_{ik} - X_{ik}\beta_k)^\top(y_{ik} - X_{ik}\beta_k - 2Z_{ik}\mathbb{E}[b_{ik}])\\
&\qquad\qquad + \text{trace}\left(Z_{ik}^\top Z_{ik}\mathbb{E}[b_{ik}b_{ik}^\top]\right)\Big]\Bigg\},\\
S(D^{-1}) &= \frac{n}{2}\left\{2D - \text{diag}(D)\right\} - \frac{1}{2}\left[2\sum_{i=1}^{n}\mathbb{E}\left[b_i b_i^\top\right] - \text{diag}\left(\sum_{i=1}^{n}\mathbb{E}\left[b_i b_i^\top\right]\right)\right],
\end{aligned}
$$

where the update for σ_k^2 was done by first rewriting (13) as $\sum_{k=1}^{K}\log\{f(y_{ik} \,|\, b_{ik}, \theta)\}$.

The score equations for γ_v and γ_y don't have closed-form solutions. Therefore, they are updated jointly using a one-step multidimensional Newton-Raphson algorithm iteration. We can write the score equation for $\gamma = \left(\gamma_v^\top, \gamma_y^\top\right)^\top$ as

$$
\begin{aligned}
S(\gamma) &= \sum_{i=1}^{n}\left[\delta_i\mathbb{E}\left[\tilde{v}_i(T_i)\right] - \int_0^{T_i}\lambda_0(u)\mathbb{E}\left[\tilde{v}_i(u)\exp\{\tilde{v}_i^\top(u)\gamma\}\right] du\right]\\
&= \sum_{i=1}^{n}\left[\delta_i\mathbb{E}\left[\tilde{v}_i(T_i)\right] - \sum_{j=1}^{J}\lambda_0(t_j)\mathbb{E}\left[\tilde{v}_i(t_j)\exp\{\tilde{v}_i(t_j)^\top\gamma\}\right] I(T_i \geq t_j)\right],
\end{aligned}
$$

where $\tilde{v}_i(t) = \left(v_i^\top, z_{i1}^\top(t)b_{i1}, \ldots, z_{iK}^\top(t)b_{iK}\right)$ is a $(q + K)$-vector, and the integration over the survival process has been replaced with a finite summation over the process evaluated at the unique failure times, since the non-parametric estimator of baseline hazard is zero except at observed failure times [1]. As $\lambda_0(t_j)$ is a function of γ, this is not a closed-form solution. Substituting $\lambda_0(t)$ by $\hat{\lambda}_0(t)$ from (6), which is a function of γ and the observed data itself, gives a score that is independent of $\lambda_0(t)$. Discussion of this in the context of univariate joint modelling is given by Hsieh et al. [14]. A useful result is that the maximum profile likelihood estimator is the same as the maximum partial likelihood estimator [18].

The observed information matrix for γ is calculated by taking the partial derivative of the score above (with $\lambda_0(t)$ substituted by $\hat{\lambda}_0(t_j)$ defined by (6)), and is given by

$$I(\gamma) \equiv -\frac{\partial}{\partial \gamma} S(\gamma) = \sum_{i=1}^{n} \sum_{j=1}^{J} \left\{ \hat{\lambda}_0(t_j) I(T_i \geq t_j) \mathbb{E}\left[\tilde{v}_i(t_j) \tilde{v}_i^\top(t_j) \exp\{\tilde{v}_i^\top(t_j)\gamma\} \right] \right\} - \sum_{j=1}^{J} \frac{\hat{\lambda}_0(t_j)^2 \Gamma(t_j)}{\sum_{i=1}^{n} \delta_i I(T_i = t_j)}.$$

where

$$\Gamma(t_j) = \left\{ \sum_{i=1}^{n} \mathbb{E}\left[\tilde{v}_i(t_j) \exp\{\tilde{v}_i^\top(t_j)\gamma\} \right] I(T_i \geq t_j) \right\} \left\{ \sum_{i=1}^{n} \mathbb{E}\left[\tilde{v}_i(t_j) \exp\{\tilde{v}_i(t_j)\gamma\} \right] I(T_i \geq t_j) \right\}^\top,$$

and $\hat{\lambda}_0(t)$ is given by (6), which is also a function of γ. In practice, calculation of $I(\gamma)$ is a computational bottleneck. Therefore, in some situations we may want to approximate it. One approximation we consider is a Gauss-Newton-like approximation, which is similar to the empirical information matrix, as defined by (12). Hence, the one-step block update at the $(m+1)$-th EM algorithm iteration is

$$\hat{\gamma}^{(m+1)} = \hat{\gamma}^{(m)} + I\left(\hat{\gamma}^{(m)}\right)^{-1} S\left(\hat{\gamma}^{(m)}\right).$$