Super Learning in Joint Models
Second report: using SL to overcome the problem of the difficulty fitting multivariate Joint Models

Arnau García

October 21, 2024

In this report we focus on the next specific framework: we have data with time to event data, as well as longitudinal data. In addition, we have more than one longitudinal variable. In particular, let us assume that we have L longitudinal biomarkers. Our main goal is to compute subject-specific dynamic predictions for survival outcomes. Under the Bayesian paradigm, which is the one we use in this work, the computation of that dynamic prediction is derived by means of the posterior distribution of the joint model. Nonetheless, this joint model is a multivariate joint model with L longitudinal outcomes. In practice, fitting a multivariate joint model is computationally expensive. In fact, for $L > 10$ it seems to be unfeasible. Taking into account that our main objective is to compute a predictive metric, we wonder if we can use super learning (SL) to calculate this metric with good enough accuracy, and avoid fitting a multivariate joint model.

This report will consist of two parts. Firstly, a mathematical introduction to the problem will be added. The problem we are dealing with will be faced using the proper mathematical context, as well as the necessary notation. In this first part our aim is to fully understand the problem, and to deal with it as rigorously as possible. In the second par of the report, a first simulation on the problem will be added as well as a case study will be analyzed. By means of the *JMbayes2* package [2] a multivariate joint model will be fitted, and the SL approach will be used to see whether the SL predictor gives similar results compared with the multivariate model.

From now on, we will be using the notation in [3], adding the detail (also treated in the previous reference) that we have L longitudinal biomarkers. Assume we have the data $\mathcal{D}_n = \{T_i, \delta_i, \boldsymbol{y}_{il}; i = 1, \ldots, n, l = 1, \ldots, L\}$. Where T_i^* denotes true time to true event of interest ε for the i-th subject, C_i the censoring time, $T_i = min(T_i^*, C_i)$ is the corresponding observed time to the event, and $\delta_i = \mathbf{1}(T_i^* \leq C_i)$ is the event indicator. Moreover, $\boldsymbol{y}_{il}$

is the $n_{il} \times 1$ longitudinal response vector for the i-th subject and the l-th longitudinal response, with element $y_{il}(t_{ij})$ being the value of the longitudinal outcome taken at time point t_{ij}, $j = 1, \ldots, n_{il}$ (n_{il} is the number of measurements for the i-th subject, l-th longitudinal response). We assume that the response vector $\boldsymbol{y}_i$ conditional on the random effects $\boldsymbol{b}_i$ has distribution $\mathcal{F}_\Psi$ within the exponential family of distributions. Thus, we are assuming a less general framework than the one postulated in [3], this is because we are interested in write the likelihood of the model, which is easier taking the assumption that the longitudinal outcomes are generalized linear mixed models, (GLMM's). The mean of the distribution of the longitudinal outcomes conditional on the random effects has the form

$$g_l\left[E(y_{il}(t)|\boldsymbol{b}_{il})\right] = m_{il}(t) = \boldsymbol{x}_{il}^\top(t)\boldsymbol{\beta}_l + \boldsymbol{z}_{il}^\top(t)\boldsymbol{b}_{il}, \quad l = 1, \ldots, L, \tag{1}$$

where $g_l(\cdot)$ denotes a known one-to-one monotonic link function for the lth longitudinal outcome, and $y_{il}(t)$ denotes the value of the lth longitudinal outcome for the ith subject at time point t, $\boldsymbol{x}_{il}(t)$ and $\boldsymbol{z}_{il}(t)$ denote the time-dependent design vectors for the fixed-effects $\boldsymbol{\beta}$ and for the random effects $\boldsymbol{b}_{il}$, respectively. We let ϕ denote the scale parameter of $\mathcal{F}_\Psi$, so $\boldsymbol{\Psi} = (\boldsymbol{\beta}, \phi)$. Notice the dependence of the previous components on the longitudinal outcome, we are assuming that we have L outcomes. The random effects are assumed to follow a multivariate normal distribution with mean zero and variance-covariance matrix D_l.

For the survival process, we assume that the risk of an event depends on a function of the subject-specific linear predictor $m_{il}(y)$ and the random effects. More specifically, we have

$$h_i\left(t \mid \mathcal{Y}_i(t), \boldsymbol{w}_i\right) = h_0(t)\exp\left\{\boldsymbol{\gamma}^\top\boldsymbol{w}_i + \sum_{l=1}^{L}\boldsymbol{\alpha}_l f_l(t, \mathcal{Y}_{il}(t), \boldsymbol{b}_{il})\right\}, \quad t > 0, \tag{2}$$

where $\mathcal{Y}_{il}(t) = \{m_{il}(s), 0 \leq s < t\}$ denotes the history of the underlying lth longitudinal process up to t, $h_o(\cdot)$ denotes the baseline hazard function, $\boldsymbol{w}_i$ is a vector of baseline covariates with corresponding regression coefficients $\boldsymbol{\gamma}$. The reader can observe that the expression above is similar to the typical one for joint models with just one longitudinal outcome, but expanded to accommodate for all longitudinal outcomes. Since we are working under the Bayesian approach, we should complete the specification of the joint model and choose how to model the baseline hazard function. We use a B-splines approach,

$$\log h_0(t) = \gamma_{h_0,0} + \sum_{q=1}^{Q}\gamma_{h_0,q}B_q(t, \boldsymbol{v}), \tag{3}$$

where $B_q(t, \boldsymbol{v})$ denotes the q-th basis function of a B-spline with knots $v_1, \ldots, v_Q$ and γ_{h_0} the vector of spline coefficients.

Since we are working under the Bayesian paradigm, we have to specify the prior distribution for all the parameters involved in our model. Let

$$\boldsymbol{\theta} = \{\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_L, \phi_1, \dots, \phi_L, \boldsymbol{\gamma}_{h_0}, \boldsymbol{\gamma}, \boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_L, vech(\boldsymbol{D}_1), \dots, vech(\boldsymbol{D}_L), \tau\} \tag{4}$$

denote the vector of all model parameters, where $vech(\boldsymbol{D}_l), l = 1, \dots, L$ denotes the unique elements of the variance-covariance matrices $\boldsymbol{D}_1, \dots, \boldsymbol{D}_L$. Inferences are drawn using the joint posterior distribution $\{\boldsymbol{\theta}, \boldsymbol{b}|\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{\delta}\}$. Standard priors will be used for $\boldsymbol{\theta}$, normal priors for all the regression coefficients, and inverse-gamma priors for $\phi_1, \dots, \phi_L$ and the diagonal elements of $\boldsymbol{D}_1, \dots, \boldsymbol{D}_L$, and LKJ prior for the correlation matrices of the random effects. To ensure smoothness of the baseline hazard function $h_0(t)$ we use a penalized prior distribution for the regression coefficients $\boldsymbol{\gamma}_{h_0}$,

$$p(\boldsymbol{\gamma}_{h_0}|\tau) \propto \tau^{\rho(K)/2} \exp\big(-\frac{\tau}{2}\boldsymbol{\gamma}_{h_0}^\top \boldsymbol{K} \boldsymbol{\gamma}_{h_0}\big),$$

where τ is the smoothing parameter that takes a Gamma$(5, 0.05)$ hyper-prior to ensure a proper posterior for $\boldsymbol{\gamma}_{h_0}$, $\boldsymbol{K} = \Delta_r^\top \Delta_r$, Δ_r denoting the rth difference penalty matrix, and $\rho(\boldsymbol{K})$ denotes the rank of $\boldsymbol{K}$. Observe that the smoothing paramter τ is also included in the parameters of the model $\boldsymbol{\theta}$. Markov Chain Monte Carlo (MCMC) is the method used to obtain samples from the posterior distribution for all model parameters and the random effects.

Having introduced the problem in a proper manner, we should remember that our main goal is to compute the following dynamic individualized prediction

$$\begin{aligned}
\pi_j(u \mid t) &= \int \mathrm{P}\left(T_j^* \geq u \mid T_j^* > t, \mathcal{Y}_j(t), \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta} \mid \mathcal{D}_n\right) d\boldsymbol{\theta} \\
&= \int \left(\int \mathrm{P}\left(T_j^* \geq u \mid T_j^* > t, \boldsymbol{b}_j, \boldsymbol{\theta}\right) p\left(\boldsymbol{b}_j \mid T_j^* > t, \mathcal{Y}_j(t), \boldsymbol{\theta}\right) d\boldsymbol{b}_j\right) p\left(\boldsymbol{\theta} \mid \mathcal{D}_n\right) d\boldsymbol{\theta} \\
&= \int\int \frac{S_j\left\{u \mid \mathcal{Y}_j\left(u, \boldsymbol{b}_j\right), \boldsymbol{\theta}\right\}}{S_j\left\{t \mid \mathcal{Y}_j\left(t, \boldsymbol{b}_j\right), \boldsymbol{\theta}\right\}} p\left(\boldsymbol{b}_j \mid T_j^* > t, \mathcal{Y}_j(t), \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta} \mid \mathcal{D}_n\right) d\boldsymbol{\theta} d\boldsymbol{b}_j.
\end{aligned} \tag{5}$$

Clearly, the quantity above is computed using the posterior distribution of the model. Due to the fact that this posterior is coming from the multivariate joint model (2), where all the L longitudinal outcomes are used, and that this model is not easy to fit, we are interested in another way of calculating this metric. Super Learning (SL) [4] is the chosen method to compute the prediction avoiding to fit the multivariate joint model.

Let us assume that we have the library $\mathcal{L} = \{M_1, \dots, M_L\}$ consisting of the following L Bayesian models:

$$\begin{aligned}
M1: &\quad h_i\left(t \mid \mathcal{Y}_{i1}(t), \boldsymbol{w}_i\right) = h_0(t) \exp\left\{\boldsymbol{\gamma}^\top \boldsymbol{w}_i + \alpha_1 f_1(t, \mathcal{Y}_{i1}(t), \boldsymbol{b}_{i1})\right\},\\
M2: &\quad h_i\left(t \mid \mathcal{Y}_{i2}(t), \boldsymbol{w}_i\right) = h_0(t) \exp\left\{\boldsymbol{\gamma}^\top \boldsymbol{w}_i + \alpha_2 f_2(t, \mathcal{Y}_{i2}(t), \boldsymbol{b}_{i2})\right\},\\
\dots &\\
M_L: &\quad h_i\left(t \mid \mathcal{Y}_{iL}(t), \boldsymbol{w}_i\right) = h_0(t) \exp\left(\boldsymbol{\gamma}^\top \boldsymbol{w}_i + \alpha_L f_L(t, \mathcal{Y}_{iL}(t), \boldsymbol{b}_{iL})\right).
\end{aligned} \tag{6}$$

As the reader can observe, the library of models above is built by decomposing the multivariate model (2) into L univariate joint model, each one with one of the L longitudinal outcomes. Another important observation is that we assume fixed $\boldsymbol{\gamma}, \boldsymbol{w}_i$ in all the L models. This is, those components of the model not related with the longitudinal outcomes (and then related with the survival outcome) are assumed to be the same in all the univariate models in $\mathcal{L}$.

The main goal of this work is to see if applying the SL procedure to the library $\mathcal{L}$ one can obtain a good enough prediction to be used instead of the multivariate model. With this objective in mind, first of all, we will try to establish a relationship between the posterior of the multivariate joint model (2) and the posteriors of the univariate models (6).

On the relationship between the posteriors

In this section we will simplify some parts of our problem in order to write as detailed as possible the posteriors of the models. When working under the Bayesian framework, the common used strategy to establish relationships between posteriors is by comparing them. It is well-known that the posterior distribution is the compromise between the likelihood and the prior distribution. Let us assume prior independence, this is, the prior distribution is the product of the corresponding marginal prior distributions. It is,

$$\begin{aligned}
p(\boldsymbol{\theta}) =& p(\boldsymbol{\beta}_1) \cdots p(\boldsymbol{\beta}_L) p(\phi_1) \cdots p(\phi_L) p(\boldsymbol{\alpha}_1) \cdots p(\boldsymbol{\alpha}_L) p(vech(\boldsymbol{D}_1)) \cdots p(vech(\boldsymbol{D}_L))\\
&\times p(\tau) p(\boldsymbol{\gamma}_{h_0}) p(\boldsymbol{\gamma}).
\end{aligned} \tag{7}$$

To specify the likelihood of the multivariate model (2), some assumptions have to be made. Besides the conditional independence between the longitudinal and the event outcomes, and the repeated measurements of the longitudinal process, on the random effects $\boldsymbol{b}_i$, when working with more than one longitudinal outcome we should add the conditional independence of the L longitudinal outcomes based, again, on the random effects [1]. Then, we assume that

$$\begin{aligned} p\left(\boldsymbol{y}_i, T_i, \delta_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}\right) &= p\left(\boldsymbol{y}_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}\right) p\left(T_i, \delta_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}\right) \\ p\left(\boldsymbol{y}_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}\right) &= \prod_l p\left(y_{il} \mid \boldsymbol{b}_{il}, \boldsymbol{\theta}_l\right), \\ p(\boldsymbol{y}_{il} | \boldsymbol{b}_{il}, \boldsymbol{\theta}_l) &= \prod_j p\left(y_{il}(t_{ij}) | \boldsymbol{b}_{il}, \boldsymbol{\theta}_l\right). \end{aligned} \tag{8}$$

Where $\boldsymbol{y}_i = (\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{iL}), \boldsymbol{y}_{il} = (y_{il}(t_{i1}), \ldots, y_{il}(t_{in_{il}}))$ and $\boldsymbol{b}_i = (\boldsymbol{b}_{i1}, \ldots, \boldsymbol{b}_{iL})$. Having done this assumption, one can easily write the likelihood function of the multivariate joint model.

In order to write the likelihood, let us assume that we have right-censored data and non-informative censoring. In addition, to simplify the algebra, we assume that the functional forms we are using in the joint models are the simple ones: $f_l(t, \mathcal{Y}_{il}(t), \boldsymbol{b}_{il}) = m_{il}(t)$. Now, using that the survival outcome contribution in the likelihood under right-censoring is

$$L = \prod_i h(t_i^*)^{\delta_i} S(t_i^*).$$

And using the assumptions exposed in (8), we have that the likelihood of the multivariate joint model (2) is

$$\begin{aligned} & p\left(\boldsymbol{y}_i, T_i, \delta_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}\right) = p\left(\boldsymbol{y}_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}\right) p\left(T_i, \delta_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}\right) \\ & = \big[\prod_{l=1}^{L} \prod_{j=1}^{n_{il}} p\left(y_{il}(t_{ij}) | \boldsymbol{b}_{il}, \boldsymbol{\theta}\right) p(\boldsymbol{b}_{il} | \boldsymbol{\theta})\big] p(T_i, \delta_i | \boldsymbol{b}_i, \boldsymbol{\theta}) \\ & \propto \exp\left\{ \sum_{l=1}^{L} \left(\frac{\sum_{j=1}^{n_{il}} (y_{il}(t_{ij}) m_{il}(t_{ij}) - A(m_{il}(t_{ij})))}{\phi_l} \right) + \sum_l \sum_j C(y_{il}(t_{ij}), \phi_l) \right\} \\ & \quad \times \prod_{l=1}^{L} \det(\boldsymbol{D}_l)^{-1/2} \exp\left(- \sum_{l=1}^{L} \boldsymbol{b}_{il}^\top \boldsymbol{D}_l^{-1} \boldsymbol{b}_{il} / 2 \right) \\ & \quad \times \left[\exp\left\{ \sum_q \gamma_{h_0,q} B_q\left(T_i, \boldsymbol{v}\right) + \boldsymbol{\gamma}^\top \boldsymbol{w}_i + \sum_{l=1}^{L} \alpha_l m_{il}\left(T_i\right) \right\} \right]^{\delta_i} \\ & \quad \times \exp\left[- \exp\left(\boldsymbol{\gamma}^\top \boldsymbol{w}_i\right) \int_0^{T_i} \exp\left\{ \sum_q \gamma_{h_0,q} B_q(s, \boldsymbol{v}) + \sum_{l=1}^{L} \alpha_l m_{il}(s) \right\} ds \right]. \end{aligned} \tag{9}$$

Now, let us write the likelihood for the lth univariant model in (6). In this case, since there is only one longitudinal outcome, we only have to take into account the conditional independence between the longitudinal and the event outcomes, and the repeated measurements of the longitudinal process, both based on the random effects.

$$
\begin{aligned}
p\left(\boldsymbol{y}_{il}, T_i, \delta_i \mid \boldsymbol{\theta}, \boldsymbol{b}_i\right) &= \prod_{j=1}^{n_{il}} p\left(y_{il}(t_{ij}) | \boldsymbol{b}_{il}, \boldsymbol{\theta}\right) p(\boldsymbol{b}_{il}|\boldsymbol{\theta}) p(T_i, \delta_i | \boldsymbol{b}_i, \boldsymbol{\theta}) \\
&\propto \exp\left\{ \frac{\sum_{j=1}^{n_{il}} (y_{il}(t_{ij}) m_{il}(t_{ij}) - A(m_{il}(t_{ij})))}{\phi_l} + \sum_{j=1}^{n_{il}} C(y_{il}(t_{ij}), \phi_l) \right\} \\
&\times \det(\boldsymbol{D}_l)^{-1/2} \exp\left(-\boldsymbol{b}_{il}^\top \boldsymbol{D}_l^{-1} \boldsymbol{b}_{il}/2\right) \\
&\times \left[\exp\left\{ \sum_q \gamma_{h_0,q} B_q\left(T_i, \boldsymbol{v}\right) + \boldsymbol{\gamma}^\top \boldsymbol{w}_i + \alpha_l m_{il}\left(T_i\right) \right\} \right]^{\delta_i} \\
&\times \exp\left[-\exp\left(\boldsymbol{\gamma}^\top \boldsymbol{w}_i\right) \int_0^{T_i} \exp\left\{ \sum_q \gamma_{h_0,q} B_q(s, \boldsymbol{v}) + \alpha_l m_{il}(s) \right\} ds \right].
\end{aligned}
\tag{10}
$$

The reader can identify in (9) and (10) the following components: in the first line after the proportionality symbol, we have the contribution of the longitudinal outcomes (or outcome). Having assumed that we model the longitudinal processes as GLMM's, one can use the general expression of the exponential family of distributions to write down the contribution of the longitudinal outcomes to the likelihood. Following, in the second line we have the contribution of the random effects. Then, in the third and forth line the reader can observe the contribution of the event outcomes.

Now, we wonder if in any way the information provided by the posteriors of the univariate models resembles the information provided by the posterior of the multivariate model. To study this, we will see if we can find any relationship between the likelihoods. The reader can observe that by making the product of the likelihoods of the L univariate models we obtain:

$$
\begin{aligned}
&\prod_{l=1}^{L} p\left(\boldsymbol{y}_{il}, T_i, \delta_i \mid \boldsymbol{\theta}, \boldsymbol{b}_i\right) \\
&\propto \exp\left\{\sum_{l=1}^{L}\left(\frac{\sum_{j=1}^{n_{il}}(y_{il}(t_{ij})m_{il}(t_{ij}) - A(m_{il}(t_{ij})))}{\phi_l}\right) + \sum_l \sum_j C(y_{il}(t_{ij}), \phi_l)\right\} \\
&\quad \times \prod_{l=1}^{L} \det(\boldsymbol{D}_l)^{-1/2} \exp\left(-\sum_{l=1}^{L} \boldsymbol{b}_{il}^{\top} \boldsymbol{D}_l^{-1} \boldsymbol{b}_{il}/2\right) \qquad (11) \\
&\quad \times \left[\exp\left\{L \sum_q \gamma_{h_0,q} B_q\left(T_i, \boldsymbol{v}\right) + L\boldsymbol{\gamma}^{\top}\boldsymbol{w}_i + \sum_{l=1}^{L} \alpha_l m_{il}\left(T_i\right)\right\}\right]^{\delta_i} \\
&\quad \times \exp\left[-\exp\left(\boldsymbol{\gamma}^{\top}\boldsymbol{w}_i\right) \int_0^{T_i} \exp\left(\sum_q \gamma_{h_0,q} B_q(s, \boldsymbol{v})\right)\left[\sum_{l=1}^{L} \exp\left(\alpha_l m_{il}(s)\right)\right] ds\right].
\end{aligned}
$$

As the reader can easily see on the expression above, the contribution of the longitudinal outcomes in the product of the likelihoods of the univariate models, is the same than in the likelihood of the multivariate model. The same happens with the contribution of the random effects. However, this is not the case for the contribution of the event outcomes. Let us analyze the differences between the contributions of the survival outcomes in (9) and (11). The first difference is the appearance of two scalars L multiplying the terms related with the baseline hazard and the regression coefficients of the survival model. These do not appear in (9), and point out that as L grows, the contribution of these terms in the likelihood is greater. In addition, while in (9) we have the sum of the longitudinal models contribution inside the exponential, in (11) we have the sum of the exponentials of the longitudinal models contribution. This difference is affecting the contribution of the survival outcome in the likelihood. We should analyze how it is actually affecting.

Let us study now the differences between the priors and how they affect in the posterior. The prior distribution for the multivariate model is (7), while the prior for the lth model is

$$p(\boldsymbol{\theta}^{(l)}) = p(\boldsymbol{\beta}_l)p(\phi_l)p(\boldsymbol{\alpha}_l)p(vech(\boldsymbol{D}_l))p(\tau)p(\boldsymbol{\gamma}_{h_0})p(\boldsymbol{\gamma}). \qquad (12)$$

Thus, when doing the product of the priors of the L models, we obtain the same prior than in the multivariate case for those parameters depending on the longitudinal outcome, but for the parameters involved in the survival process we obtain $p(\tau)^L p(\boldsymbol{\gamma}_{h_0})^L p(\boldsymbol{\gamma})^L$. This makes sense since these terms are repeated in all the L univariate models. In a way, this tells us that when L grows, the contribution of these priors in the posterior grows as well. This is logical, since by pooling the information from the L multivariate models we are using L times the information provided by these parameters.

Mayo Clinic Primary Biliary Cirrhosis Data analysis

The *Mayo Clinic Primary Biliary Cirrhosis Data* (`pbc2`) is a data set available with the *JMbayes2* package. In this data set we have the follow up of 312 randomised patients with primary biliary cirrhosis, a rare autoimmune liver disease, at Mayo Clinic [1]. It is not the purpose of this report to go into the details of this database. However, this database has been chosen to see a practical case of the problem we are dealing with since it presents several longitudinal processes. In this first case study, the main objective has been to become familiar with the functions of the *JMbayes2* package that allow us to deal with the objects of interest for our work. Moreover, the objective of this first simple case study was to see that we are on the right track with our hypothesis and that Super Learning allows us to compute predictions with a good accuracy, so that we avoid fitting a multivariate model. This case study is based in the one exposed in [3].

We consider five different longitudinal outcomes, and we use the following longitudinal submodels for each one:

$$\begin{aligned}
\log(\texttt{serBilir}(t_{ij})) &= m_{i1}(t_{ij}) + \varepsilon_{i1}(t_{ij}) \\
&= (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_{ij} + \varepsilon_{i1}(t_{ij}), \\
\texttt{prothrombin}(t_{ij}) &= m_{i2}(t_{ij}) + \varepsilon_{i2}(t_{ij}) \\
&= (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_{ij} + \beta_2\texttt{sex}_i + \beta_3\texttt{sex}_i * t_{ij} + \varepsilon_{i2}(t_{ij}), \\
\log\left(\frac{p(\texttt{ascites}(t_{ij}) = 1)}{1 - p(\texttt{ascites}(t_{ij}) = 1)}\right) &= m_{i3}(t_{ij}) + \varepsilon_{i3}(t_{ij}) \\
&= (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_{ij} + \beta_2\texttt{sex}_i + \varepsilon_{i3}(t_{ij}), \\
\texttt{albumin}(t_{ij}) &= m_{i4}(t_{ij}) + \varepsilon_{i4}(t_{ij}) \\
&= (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_{ij} + \varepsilon_{i4}(t_{ij}), \\
\log\left(\frac{p(\texttt{hepatomegaly}(t_{ij}) = 1)}{1 - p(\texttt{hepatomegaly}(t_{ij}) = 1)}\right) &= m_{i5}(t_{ij}) + \varepsilon_{i5}(t_{ij}) \\
&= (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_{ij} + \varepsilon_{i5}(t_{ij}).
\end{aligned}$$

Where time is in years. Then, we have three linear mixed models (LMM) and two generalized linear mixed models (GLMM). In the latter random intercepts and random slopes are assumed uncorrelated. The longitudinal processes above are used to fit a multivariate joint model

[1] See further information about the data set in `https://cran.r-project.org/web/packages/JMbayes2/JMbayes2.pdf`

(with $L = 5$ different longitudinal outcomes). In this multivariate joint model, we have for the survival process

$$h_i\left(t \mid \mathcal{Y}_i(t), \boldsymbol{w}_i\right) = h_0(t) \exp\left\{\gamma \texttt{sex}_i + \sum_{l=1}^{5} \alpha_l m_{il}(t)\right\}, \quad t > 0.$$

We have fitted the joint model and obtained the posterior distribution by means of MCMC (implemented in *JMbayes2*). We have used the integrated Brier score (IBS) and the expected predictive cross-entropy (EPCE) as predictive performance metrics. We evaluate the predictive performance of the multivariate model in two time intervals $(t, t+\Delta t]$, for instance $(5, 8]$ and $(7, 10]$. 202 individuals were at risk at year 5, besides 28 events, 70 censored times occurred in the interval $(5, 8]$. Moreover, 129 individuals were at risk at time 7, 22 events occurred in $(7, 10]$ and there were 56 censored times in that interval. The results obtained for the multivariate model in $(5, 8]$ were: $\text{IBS} = 0.0575, \text{EPCE} = 0.4419$. In addition, the results for $(7, 10]$ are: $\text{IBS} = 0.0729, \text{EPCE} = 0.6112$.

Super Learning procedure has been applied using the library built by the following $L = 5$ univariate joint models:

$$\begin{aligned}
M1: &\quad h_i\left(t \mid \mathcal{Y}_{i1}(t), \texttt{sex}_i\right) = h_0(t) \exp\left\{\gamma \texttt{sex}_i + \alpha_1 m_{i1}(t)\right\},\\
M2: &\quad h_i\left(t \mid \mathcal{Y}_{i2}(t), \texttt{sex}_i\right) = h_0(t) \exp\left\{\gamma \texttt{sex}_i + \alpha_2 m_{i2}(t)\right\},\\
M3: &\quad h_i\left(t \mid \mathcal{Y}_{i3}(t), \texttt{sex}_i\right) = h_0(t) \exp\left\{\gamma \texttt{sex}_i + \alpha_3 m_{i3}(t)\right\},\\
M4: &\quad h_i\left(t \mid \mathcal{Y}_{i4}(t), \texttt{sex}_i\right) = h_0(t) \exp\left\{\gamma \texttt{sex}_i + \alpha_4 m_{i4}(t)\right\},\\
M5: &\quad h_i\left(t \mid \mathcal{Y}_{i5}(t), \texttt{sex}_i\right) = h_0(t) \exp\left\{\gamma \texttt{sex}_i + \alpha_5 m_{i5}(t)\right\}.
\end{aligned}$$

We have split the data into three folds and fitted the five univariate joint models. Taking profit of the functions available in the *JMbayes2* package, the Super Learning procedure has been applied and we have obtained the results exposed in Table 1 and Table 2. In those tables *SL (full)* is referring to the accuracy of the model using the weights computed by the SL procedure, but trained and tested with the whole data set.

Notice that, since IBS and EPCE are two scoring rules [3], the lower score indicates the better accuracy. Thus, in the results presented, those models with lower IBS and EPCE are the ones with better accuracy. We observed that for both time intervals the differences between the integrated Brier score (IBS) are small. In both time intervals, the model with smaller IBS is M_4. Moreover, M_4 is the dominating model in the SL procedure in both time intervals in terms of weight. In both time intervals the super learning estimates of the IBS are smaller than the univariate joint model estimates (with small differences with M_4), but they are bigger than the multivariate joint model IBS estimates. This is, SL procedure is improving

	$(5, 8]$		$(7, 10]$	
	IBS	Weights	IBS	Weights
Multi	0.0575		0.0729	
SL	0.0645		0.0758	
SL (full)	0.0642		0.0734	
M_1	0.0737	0.0088	0.0872	0.3039
M_2	0.0744	0.0084	0.0889	0
M_3	0.0922	$1 \cdot e^{-4}$	0.1101	0.0012
M_4	0.0646	0.9825	0.0781	0.689
M_5	0.0869	$2 \cdot e^{-4}$	0.0904	0.0059

Table 1: IBS results for the SL procedure, the multivariate model, and the SL weights applied to the whole data set. The SL procedure is based in 3-fold cross-validation.

	$(5, 8]$		$(7, 10]$	
	EPCE	Weights	EPCE	Weights
Multi	0.4419		0.6112	
SL	0.3703		0.4645	
SL (full)	0.3735		0.4365	
M_1	0.4801	0.3884	0.6795	0.0964
M_2	0.5164	0.002	0.607	0.0037
M_3	0.4826	0.2602	0.5632	0.2964
M_4	0.5035	0.2813	0.6172	0.2134
M_5	0.4994	0.0681	0.5035	0.3902

Table 2: EPCE results for the SL procedure, the multivariate model, and the SL weights applied to the whole data set. The SL procedure is based in 3-fold cross-validation.

the predictive performance of the univariate models separately (with no big differences with respect M_4), but the multivariate joint model still has better predictive perfomence.

If we analyze the results on the expected predictive cross-entropy (see Table 2), we can see that for both time intervals the super learning estimates of the EPCE are smaller than in each of the five univariate models. In addition, the super learning estimates of the EPCE are smaller than in the multivariate models, for both intervals. Furthermore, the reader can observe in Table 2 that the differences for the EPCE are bigger than for the IBS. This time, the dominating models in terms of weights for the SL procedure are M_1 and M_4, for the time intervals $(5, 8]$ and $(7, 10]$. Nonetheless, the differences between weights are not as big as in the case of the IBS.

Simulation study

Based on the simulation study presented in [3], as well as in the supplementary material of the previous paper, a simulation study has been carried out to evaluate the performance of the SL algorithm and to assess whether SL can be useful to compute predictions and avoid the use of a multivariate joint model. To code the simulation I have used `https://drizopoulos.github.io/JMbayes2/articles/Non_Gaussian_Mixed_Models.html#negative-binomial-mixed-models` as a benchmark.

For the longitudinal outcomes, we consider $L = 5$ longitudinal processes. Three of them will be linear mixed models, and the remaining two models will be generalized linear mixed models (see model specification below). For the survival outcome, we consider the hazard model of the multivariate joint model (see model specification below). For this first simulation, two scenarios has been planned. Future simulations will explore adding more scenarios that can add robustness to the results. In the first scenario one data-generating model has been considered, a multivariate joint model. In the second scenario, we use six data-generating models: the multivariate joint model, and the $L = 5$ univariate joint models. We have simulated training and testing data, both contain 300 subjects. In the second scenario one-sixth of the subjects are generated by the multivariate model, one-sixth by the first univariate model, and so on. We have simulated a total amount of 20 data sets in the first scenario (the objective was to generate more datasets, but seeing the computational cost and that 20 datasets did not achieve the expected results, no more were generated). In the second scenario 100 data sets have been generated.

For each simulated dataset, we implemented three-fold cross-validation, fitting the $L = 5$ univariate joint models (composed for the longitudinal sub-models separated) and the multivariate joint model. The cross-validated predictions from each model were used to estimate the model weights to optimize predictive performance via the super learning procedure. The discrete super learner was also compared with the ensemble super learner. Integrated Brier score and EPCE are the chosen predictive performance measures. We have assessed the predictive performance in the interval $(t, t + \Delta] = (12, 18]$. Let us expose the model specification and the parameter values used, the longitudinal models used are:

$$\begin{aligned}
y_{i1}(t_{ij}) &= m_{i1}(t_{ij}) + \varepsilon_{i1}(t_{ij}) \\
&= (\beta_0^1 + b_{0i}^1) + (\beta_1^1 + b_{1i})t_{ij} + \beta_2^1 \texttt{sex}_i + \beta_3^1 \texttt{sex}_i * t_{ij} + \varepsilon_{i1}(t_{ij}), \\
y_{i2}(t_{ij}) &= m_{i2}(t_{ij}) + \varepsilon_{i2}(t_{ij}) \\
&= (\beta_0^2 + b_{0i}^2) + (\beta_1^2 + b_{1i}^2)t_{ij} + \beta_2^2 \texttt{sex}_i + \varepsilon_{i2}(t_{ij}), \\
y_{i3}(t_{ij}) &= m_{i3}(t_{ij}) + \varepsilon_{i3}(t_{ij}) \\
&= (\beta_0^3 + b_{0i}^3) + (\beta_1^3 + b_{1i}^3)t_{ij} + \varepsilon_{i3}(t_{ij}), \\
\log\left(\frac{p(y_{i4}(t_{ij}) = 1)}{1 - p(y_{i4}(t_{ij}) = 1)}\right) &= m_{i4}(t_{ij}) + \varepsilon_{i4}(t_{ij}) \\
&= (\beta_0^4 + b_{0i}^4) + (\beta_1^4 + b_{1i}^4)t_{ij} + \beta_2^4 \texttt{sex}_i + \varepsilon_{i4}(t_{ij}), \\
\log\left(\frac{p(y_{i5}(t_{ij}) = 1)}{1 - p(y_{i5}(t_{ij}) = 1)}\right) &= m_{i5}(t_{ij}) + \varepsilon_{i5}(t_{ij}) \\
&= (\beta_0^5 + b_{0i}^5) + (\beta_1^5 + b_{1i}^5)t_{ij} + \varepsilon_{i5}(t_{ij}).
\end{aligned}$$

Where the super-indices in the parameters denote the number of the longitudinal model. As the reader can see, we have defined three linear mixed models, and two generalized linear mixed models. Thus, we are following the same structure as in the previous section. The values chosen for the fixed effects were $(\beta_0^1, \beta_1^1, \beta_2^1, \beta_3^1) = (-2.2, -0.25, 1.24, -0.05)$, $(\beta_0^2, \beta_1^2, \beta_2^2) = (-1.8, -0.06, 0.5)$, $(\beta_0^3, \beta_1^3) = (-2.5, 0.0333)$, $(\beta_0^4, \beta_1^4, \beta_2^4) = (0.01, 0.5, -0.31416)$ and $(\beta_0^5, \beta_1^5) = (1, 0.155)$. For the error terms of the first three models we assumed $\varepsilon_{il} \sim \mathcal{N}(0, \sigma_l^2)$, where $\sigma_1^2 = 0.125, \sigma_2^2 = 0.25, \sigma_3^2 = 0.25$. The random effects were assumed to follow a multivariate normal distribution with mean zero and the following covariance matrices:

$$\begin{aligned}
D_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0.5 \end{pmatrix}, D_2 = \begin{pmatrix} 1.2 & 0 \\ 0 & 0.25 \end{pmatrix}, D_3 = \begin{pmatrix} 0.15 & 0 \\ 0 & 0.05 \end{pmatrix}, \\
D_4 = \begin{pmatrix} 0.212 & 0 \\ 0 & 0.0125 \end{pmatrix}, D_4 = \begin{pmatrix} 0.01321 & 0 \\ 0 & 0.0754 \end{pmatrix}.
\end{aligned}$$

Then, as the reader can observe in the covariance matrices above, we assumed that random effects are mutually independent (all covariances are 0). From each model and for each subject we simulated longitudinal responses at time zero anf then at 14 randomly selected time points coming from $U(0, 20)$. For the event time outcome, in the first scenario, we simulated event times from the following hazard model by means of the inverse transform sampling method,

$$h_i\left(t \mid \mathcal{Y}_i(t), \boldsymbol{w}_i\right) = h_0(t) \exp\left\{\gamma_0 + \gamma_1 \texttt{sex}_i + \sum_{l=1}^{5} \alpha_l m_{il}(t)\right\}, \quad t > 0, \tag{13}$$

where $h_0(t) = \phi t^{\phi-1}$ (thus, we are working with the Weibull distribution for the baseline hazard, the other parameter is γ_0, presented as an intercept inside the exponential function), with $\phi = 2.5$, $\gamma_0 = -25$, $\gamma_1 = 0.5$ and $\texttt{sex}_i$ denotes a binary sex indicator. In addition, the values for the association parameters are $\alpha_1 = 0.8, \alpha_2 = 0.145, \alpha_3 = 0.61, \alpha_4 = 0.05, \alpha_5 = 0.38$. The censored mechanism used was fixed Type I, all event times greater than 25 were censored. Moreover, the longitudinal measurements that were taken after the observed time were dropped.

In the second scenario, in addition to (13), the following univariate joint models have been used:

$$\begin{aligned}
M1: &\quad h_i\left(t \mid \mathcal{Y}_{i1}(t), \texttt{sex}_i\right) = h_0(t) \exp\left\{\gamma_0 + \gamma_1 \texttt{sex}_i + \alpha_1 m_{i1}(t)\right\}, \\
M2: &\quad h_i\left(t \mid \mathcal{Y}_{i2}(t), \texttt{sex}_i\right) = h_0(t) \exp\left\{\gamma_0 + \gamma_1 \texttt{sex}_i + \alpha_2 m_{i2}(t)\right\}, \\
M3: &\quad h_i\left(t \mid \mathcal{Y}_{i3}(t), \texttt{sex}_i\right) = h_0(t) \exp\left\{\gamma_0 + \gamma_1 \texttt{sex}_i + \alpha_3 m_{i3}(t)\right\}, \\
M4: &\quad h_i\left(t \mid \mathcal{Y}_{i4}(t), \texttt{sex}_i\right) = h_0(t) \exp\left\{\gamma_0 + \gamma_1 \texttt{sex}_i + \alpha_4 m_{i4}(t)\right\}, \\
M5: &\quad h_i\left(t \mid \mathcal{Y}_{i5}(t), \texttt{sex}_i\right) = h_0(t) \exp\left\{\gamma_0 + \gamma_1 \texttt{sex}_i + \alpha_5 m_{i5}(t)\right\},
\end{aligned}$$

where the parameters and the model specification for the baseline hazard function are the same than above.

Now, let us expose the results of the simulations done. The results of the first scenario are shown in the boxplots exposed in Figure 1. As the reader can observe, the results fir those metrics computed via super learning are worse than the results for the multivariate joint model (which is the only data-generating model in this scenario). The results of the second scenario are exposed in Figure 2. As we can observe, the results seem to be more stable, and there are no big differences between the metrics computed via super learning and the ones computed via the multivariate joint model. In addition, we have obtained, on average, 34% and 40% of censoring for the first and second scenario, respectively.

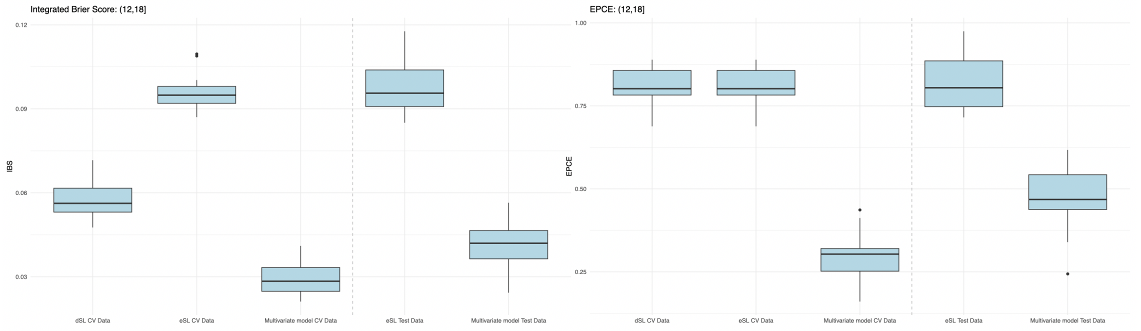

Figure 1: Simulation results for the EPCE and the IBS under Scenario I. The left panel show the results for the IBS, and the right one for the EPCE, both for the time interval $(12, 18]$. In each panel, five boxplots are shown summarizing the results over the 20 simulated datasets; in the left part of the dashed vertical lines there are metrics computed with the training data, in the right part with the test data. Each simulated training and testing dataset contains 300 subjects. Results based on 3-fold CV.

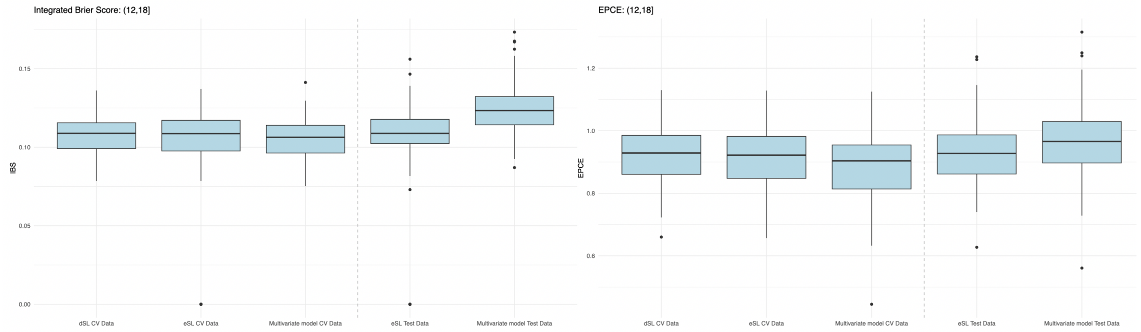

Figure 2: Simulation results for the EPCE and the IBS under Scenario II. The left panel show the results for the IBS, and the right one for the EPCE, both for the time interval $(12, 18]$. In each panel, five boxplots are shown summarizing the results over the 100 simulated datasets; in the left part of the dashed vertical lines there are metrics computed with the training data, in the right part with the test data. Each simulated training and testing dataset contains 300 subjects. Results based on 3-fold CV.

Conclusions and remarks

To conclude this report some conclusions and remarks per each section will be added.

- **On the relationship between the posteriors:** the main goal of this section is to establish a relationship between the posterior of the multivariate joint model and the L posteriors of the univariate joint models. We have seen that the posterior of these models are clearly related. Nonetheless, I would like to explore more in depth this approach. Maybe the priors have to be included explicitly to perform a more detailed analysis. Moreover, the differences between the likelihood of the multivariate joint model and the product of likelihoods of the univariate joint models, should be further analyzed. I think that, in order to see in a mathematical way that multivariate and univariate models are related, and that therefore it makes sense that super learning (which is an algorithm that makes use of the posteriors of univariate models) is useful to make predictions of multivariate models, this approach can be interesting. Analyzing what happens with posteriors within the SL algorithm itself becomes much more complicated, as cross validation and other factors come into play. But I think that seeing that posteriors can be related in general can be interesting.

- **Mayo Clinic Primary Biliary Cirrhosis Data analysis:** the purpose of this section was to conduct a small case study. In this way, I was able to familiarize myself with the tools available in the *JMbayes2* package, as well as how the results should be reported and which metrics should be reported in these cases. It is clear that these data do not necessarily have to be the ones used in the final work. The data have been useful to me to carry out this report.

- **Simulation study:** the purpose of this section has been to carry out a first simulation. This is the first time I have done simulations of this type, and carrying out this simulation has been a great help to familiarize myself with the methods used. For the moment I have only simulated one scenario, since the computational cost is high and therefore the time required for these simulations is quite high. Other scenarios are going to have to be included in the future, and also some aspects of the simulation need to be revised (i.e. using non-independent random effects, other model specifications, etc.). I want to ask some questions about the simulation.

  – First, is the percentage of censoring fixed in for each generated data set?

  – The discrete SL is chosen in each generated data set? It is, do I have to look for each data set the model with minimum score and this is the dSL? (I guess so, and this is the method I have used).

  – Is there any way to choose the parameters so that a correct censorship percentage is obtained and at the same time the events are distributed more or less uniformly over time? In order to have an adequate level of censorship I was playing with the parameters, but in my simulations the events happen for quite long times in general.

  – When simulating the datasets, should I use the multivariate model as a data generating model? Or, should I use the multivariate joint model, as well as the different univariate joint models to generate the data? I think it makes sense to use the multivariate model as the oracle model, and generate the data only with this one, since it is the model we want to approximate via SL. However, I don't know if I am committing some kind of "overfitting" by generating the data only with the multivariate model and then calculating the prediction scores for all models. It seems that the results for the scenario II are more stable.

- Of course, simulations with more data sets will be carried out. For the moment in this report I have only included these two simulations as I did not want to make the submission of the report any longer. The simulations have turned out to be more computationally expensive than I expected. As these have been preliminary simulations to familiarize myself with the procedure necessary to do a simulation, I have decided to add in the report these small simulations and leave the computer working on longer simulations in the next few days. Finally, for the second scenario, 100 datasets have been used.

References

[1] Dimitris Rizopoulos. *Joint models for longitudinal and time-to-event data: With applications in R.* CRC press, 2012.

[2] Dimitris Rizopoulos, Grigorios Papageorgiou, and P Miranda Afonso. Jmbayes2: extended joint models for longitudinal and time-to-event data. *R package. v. 0.4-5. Available online: https://drizopoulos. github. io/JMbayes2/(accessed on 26 July 2022)*, 2022.

[3] Dimitris Rizopoulos and Jeremy MG Taylor. Optimizing dynamic predictions from joint models using super learning. *Statistics in Medicine*, 43(7):1315–1328, 2024.

[4] Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.