Super Learning in Joint Models

Third report: Super learning in multivariate joint models for longitudinal and time-to-event data for optimizing dynamic predictions. A comparison with functional principal component analysis-based methods.

Arnau G. Fernández

MESIO UPC-UB

Director: Dimitris Rizopoulos

Co-Director: Guadalupe Gómez

Final Master's Thesis

Contents

1. Introduction

2. On the relationship between the posteriors

3. Simulation study

4. Comparison with functional principal component analysis-based methods
   - FPCA-based methods
   - MFPCA-Cox model
   - MFPCA-DeepSurv
   - Simulation study

5. Further comments: about BMA

Goals of third report

- Goal 1: Solve the issues on the theoretical concepts commented in the previous meeting.
- Goal 2: Perform simulation studies taking into account the considerations that were discussed at the last meeting.
- Goal 3: In order to give arguments in favor of the super learning algorithm, explore other dynamic prediction methods used in the literature and confront them with SL.

Introduction: assumptions and notation

We have the data $\mathcal{D}_n = \{T_i, \delta_i, \boldsymbol{y}_{li}; i = 1, \ldots, n, l = 1, \ldots, L\}$.

- Where T_i^* denotes true time to the 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, $\boldsymbol{y}_{li}$ is the $n_{li} \times 1$ longitudinal response vector for the i-th subject and the l-th longitudinal response.

Introduction: assumptions and notation

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. The mean is:

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

where:

- $g_l(\cdot)$ denotes a known one-to-one monotonic link function for the lth longitudinal outcome
- $\boldsymbol{x}_{li}(t)$ and $\boldsymbol{z}_{li}(t)$ denote the time-dependent design vectors for the fixed-effects β and for the random effects $\boldsymbol{b}_{li}$
- ϕ denote the scale parameter

Random effects

- To account for the association between the multiple longitudinal outcomes, we link their corresponding random effects.
- More specifically, the complete vector of random effects $\boldsymbol{b}_i = (\boldsymbol{b}_{1i}^\top, \boldsymbol{b}_{2i}^\top, \dots, \boldsymbol{b}_{Li}^\top)^\top$ is assumed to follow a multivariate normal distribution with mean zero and variance-covariance matrix $\boldsymbol{D}$.

Introduction: assumptions and notation

For the survival process, we have

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

where:

- $\mathcal{Y}_{li}(t) = \{m_{li}(s), 0 \leq s < t\}$ denotes the history of the underlying lth longitudinal process up to t
- $h_0(\cdot)$ denotes the baseline hazard function
- $\boldsymbol{w}_i$ is a vector of baseline covariates with corresponding regression coefficients γ

Introduction: assumptions and notation

We use a *B*-splines approach to specify the baseline hazard,

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

where

- $B_q(t, \boldsymbol{v})$ denotes the q-th basis function of a B-spline with knots $v_1, \dots, v_Q$,
- γ_{h_0} the vector of spline coefficients.
- Q is the number of knots we use.

Introduction: assumptions and notation

We have to specify the prior distribution for all the parameters involved in our model. All model parameters:

$$\begin{aligned}\boldsymbol{\theta} =& \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_L, \phi_1, \ldots, \phi_L, \gamma_{h_0}, \gamma, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_L, \\ & \mathit{vech}(\boldsymbol{D}), \tau\}\end{aligned}$$

where:

- $\mathit{vech}(\boldsymbol{D})$ denotes the unique elements of the variance-covariance matrix $\boldsymbol{D}$.
- Normal priors for all the regression coefficients, and inverse-gamma priors for $\phi_1, \ldots, \phi_L$ and the diagonal elements of $\boldsymbol{D}$, and LKJ prior for the correlation matrices of the random effects.

Dynamic individualized predictions

Our main goal is to compute the following dynamic individualized prediction:

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

Introduction: assumptions and notation

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

$$
\begin{aligned}
M1: \quad & h_i\left(t \mid \mathcal{Y}_{1i}(t), \boldsymbol{w}_i\right) = h_{0,1}(t) \exp\left\{\boldsymbol{\gamma_1}^\top \boldsymbol{w}_i + f_1(t, \mathcal{Y}_{1i}(t), \boldsymbol{b}_{1i}, \alpha_1)\right\}, \\
M2: \quad & h_i\left(t \mid \mathcal{Y}_{2i}(t), \boldsymbol{w}_i\right) = h_{0,2}(t) \exp\left\{\boldsymbol{\gamma_2}^\top \boldsymbol{w}_i + f_2(t, \mathcal{Y}_{2i}(t), \boldsymbol{b}_{2i}, \alpha_2)\right\}, \\
\ldots \\
M_L: \quad & h_i\left(t \mid \mathcal{Y}_{Li}(t), \boldsymbol{w}_i\right) = h_{0,L}(t) \exp\left\{\boldsymbol{\gamma_L}^\top \boldsymbol{w}_i + f_L(t, \mathcal{Y}_{Li}(t), \boldsymbol{b}_{Li}, \alpha_L)\right\}.
\end{aligned}
$$

Super Learning algorithm will be applied to the library $\mathcal{L}$.

Posteriors analysis: assumptions

Let us assume prior independence:

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

To specify the likelihood of the multivariate model, some assumptions have to be made:

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

Assumptions

- We assume we have non-informative right censored data. So the survival outcome likelihood is computed as

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

- We assume the functional forms $f_l(t, \mathcal{Y}_{li}(t), \boldsymbol{b}_{li}, \alpha_l) = \alpha_l m_{li}(t)$ for all $l = 1, \ldots, L$.

Likelihood of the multivariate JM

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

Likelihood of the product of the univariate JM's

$$\prod_{l=1}^{L} p\left(\boldsymbol{y}_{li}, T_i, \delta_i \mid \boldsymbol{\theta}, \boldsymbol{b}_{li}\right) = \prod_{l=1}^{L} \Big[\prod_{j=1}^{n_{li}} p\left(y_{li}(t_{ij}) | \boldsymbol{b}_{li}, \boldsymbol{\theta}\right) \Big] p(\boldsymbol{b}_{li}|\boldsymbol{\theta}) p(T_i, \delta_i | \boldsymbol{b}_{li}, \boldsymbol{\theta})$$

$$= \left(\prod_{l=1}^{L} \Big[\prod_{j=1}^{n_{li}} p\left(y_{li}(t_{ij}) | \boldsymbol{b}_{li}, \boldsymbol{\theta}\right) \Big] \right) \left(\prod_{l=1}^{L} p(\boldsymbol{b}_{li}|\boldsymbol{\theta}) \right) \left(\prod_{l=1}^{L} p(T_i, \delta_i | \boldsymbol{b}_{li}, \boldsymbol{\theta}) \right)$$

$$\propto \exp\left\{ \sum_{l=1}^{L} \left(\frac{\sum_{j=1}^{n_{li}} (y_{li}(t_{ij}) m_{li}(t_{ij}) - A(m_{li}(t_{ij})))}{\phi_l} \right) + \sum_{l} \sum_{j} C(y_{li}(t_{ij}), \phi_l) \right\}$$

$$\times \prod_{l=1}^{L} \det(\boldsymbol{D}_l)^{-1/2} \exp\left(-\sum_{l=1}^{L} \boldsymbol{b}_{li}^{\top} \boldsymbol{D}_l^{-1} \boldsymbol{b}_{li} / 2 \right)$$

Continuation

$$
\times \left[\exp\left\{\sum_{l=1}^{L}\left(\sum_{q}\gamma_{h_0,q,l}B_{q,l}\left(T_i,\boldsymbol{v}\right)\right)+\left(\sum_{l=1}^{L}\boldsymbol{\gamma}_l^{\top}\right)\boldsymbol{w}_i+\sum_{l=1}^{L}\alpha_l m_{li}\left(T_i\right)\right\}\right]^{\delta_i}
$$

$$
\times \exp\left[-\sum_{l=1}^{L}\exp\left(\boldsymbol{\gamma_l}^{\top}\boldsymbol{w}_i\right)\int_0^{T_i}\exp\left(\sum_{q}\gamma_{h_0,q,l}B_{q,l}(s,\boldsymbol{v})\right)\left[\sum_{l=1}^{L}\exp\left(\alpha_l m_{li}(s)\right)\right]ds\right].
$$

Comparison between likelihoods

- The contribution of the longitudinal outcomes is identical. Notice that we have

$$\Big[\prod_{l=1}^{L}\prod_{j=1}^{n_{li}} p\left(y_{li}(t_{ij})|\boldsymbol{b}_{li},\theta\right)\Big] \quad = \quad \left(\prod_{l=1}^{L}\Big[\prod_{j=1}^{n_{li}} p\left(y_{li}(t_{ij})|\boldsymbol{b}_{li},\theta\right)\Big]\right),$$

  where we have the contribution of the longitudinal outcomes to the likelihood in the multivariate JM and product of univariate JMs, on the left and the right part, respectively.

Comparison between likelihoods

- The contribution of the random effects is different. In particular, we have

$$p(\boldsymbol{b}_i|\boldsymbol{\theta}) \quad \neq \quad \left(\prod_{l=1}^{L} p(\boldsymbol{b}_{li}|\boldsymbol{\theta})\right),$$

  having the contribution for the multivariate joint model on the left part, as well as the contribution for the product of the L likelihoods on the right part.

Comparison between likelihoods

- The contribution of the survival outcomes is different. For each univariate joint model the γ_l comes from $[T_i, \delta_i | \boldsymbol{b}_{li}, \theta]$ (idem for $\gamma_{h_0,q,l}$). We can see easily the difference when comparing:

$$p(T_i, \delta_i | \boldsymbol{b}_i, \boldsymbol{\theta}) \quad \neq \quad \left(\prod_{l=1}^{L} p(T_i, \delta_i | \boldsymbol{b}_{li}, \boldsymbol{\theta}) \right),$$

  where the contribution in the multivariate JM is on the left part, and the contribution in the product of univariate JMs is on the right part.

Difference between priors

The prior for the lth univariate joint model is

$$p(\boldsymbol{\theta}_l) = p(\boldsymbol{\beta}_l)p(\phi_l)p(\boldsymbol{\alpha}_l)p(vech(\boldsymbol{D}_l))p(\tau_l)p(\gamma_{h_{0,l}})p(\gamma_l).$$

When doing the product of the priors of the L models, we obtain the following

$$\begin{aligned}\prod_{l=1}^{L} p(\boldsymbol{\theta}_l) = p(\boldsymbol{\beta}_1)\cdots p(\boldsymbol{\beta}_L)p(\phi_1)\cdots p(\phi_L)p(\boldsymbol{\alpha}_1)\cdots p(\boldsymbol{\alpha}_L)p(vech(\boldsymbol{D}_1))\cdots \\ p(vech(\boldsymbol{D}_L)) \times p(\tau_1)\cdots p(\tau_L)p(\gamma_{h_{0,1}})\cdots p(\gamma_{h_{0,L}})p(\gamma_1)\cdots p(\gamma_L) \\ \propto p(\boldsymbol{\beta}_1)\cdots p(\boldsymbol{\beta}_L)p(\phi_1)\cdots p(\phi_L)p(\boldsymbol{\alpha}_1)\cdots p(\boldsymbol{\alpha}_L) \\ \times p(vech(\boldsymbol{D}_1))\cdots p(vech(\boldsymbol{D}_L))p(\tau)^L p(\gamma_{h_0})^L p(\gamma)^L.\end{aligned}$$

Difference between posteriors

Taking into account that we are using non-informative priors, the impact of these a prior distributions on the posterior distributions is not expected to be very significant. Thus, differences between likelihoods will be much more important when comparing between posteriors.

Simulation study

The multivariate joint model is the only data-generating model. We consider $L = 5$ longitudinal processes.

- Scenario I: We have used non-independent random effects within longitudinal processes, but independent random effects between longitudinal outcomes. Random censoring has been the mechanism used in this case.
- Scenario II: We have used non-independent random effects within and between longitudinal processes, as well as Type I (administrative) censoring.

Longitudinal processes

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:

$$
\begin{aligned}
y_{1i}(t_{ij}) &= m_{1i}(t_{ij}) + \varepsilon_{1i}(t_{ij}) \\
&= (\beta_0^1 + b_{0i}^1) + (\beta_1^1 + b_{1i})t_{ij} + \beta_2^1 \mathtt{sex}_i + \beta_3^1 \mathtt{sex}_i t_{ij} + \varepsilon_{1i}(t_{ij}), \\
y_{2i}(t_{ij}) &= m_{2i}(t_{ij}) + \varepsilon_{2i}(t_{ij}) \\
&= (\beta_0^2 + b_{0i}^2) + (\beta_1^2 + b_{1i}^2)t_{ij} + \beta_2^2 \mathtt{sex}_i + \varepsilon_{2i}(t_{ij}), \\
y_{3i}(t_{ij}) &= m_{3i}(t_{ij}) + \varepsilon_{3i}(t_{ij}) \\
&= (\beta_0^3 + b_{0i}^3) + (\beta_1^3 + b_{1i}^3)t_{ij} + \varepsilon_{3i}(t_{ij}),
\end{aligned}
$$

Longitudinal processes

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

Survival outcomes

The multivariate joint model:

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

The five univariate joint models that make up the library $\mathcal{L} = \{M_1, M_2, M_3, M_4, M_5\}$:

$$\begin{aligned}
M1: &\quad h_i\left(t \mid \mathcal{Y}_{1i}(t), \texttt{sex}_i\right) = h_0(t) \exp\left\{\gamma_0 + \gamma_1 \texttt{sex}_i + \alpha_1 m_{1i}(t)\right\},\\
M2: &\quad h_i\left(t \mid \mathcal{Y}_{2i}(t), \texttt{sex}_i\right) = h_0(t) \exp\left\{\gamma_0 + \gamma_1 \texttt{sex}_i + \alpha_2 m_{2i}(t)\right\},\\
M3: &\quad h_i\left(t \mid \mathcal{Y}_{3i}(t), \texttt{sex}_i\right) = h_0(t) \exp\left\{\gamma_0 + \gamma_1 \texttt{sex}_i + \alpha_3 m_{3i}(t)\right\},\\
M4: &\quad h_i\left(t \mid \mathcal{Y}_{4i}(t), \texttt{sex}_i\right) = h_0(t) \exp\left\{\gamma_0 + \gamma_1 \texttt{sex}_i + \alpha_4 m_{4i}(t)\right\},\\
M5: &\quad h_i\left(t \mid \mathcal{Y}_{5i}(t), \texttt{sex}_i\right) = h_0(t) \exp\left\{\gamma_0 + \gamma_1 \texttt{sex}_i + \alpha_5 m_{5i}(t)\right\}.
\end{aligned}$$

Some model specifications

- Event times have been simulated by means of the inverse transform sampling method.
- The baseline hazard function is $h_0(t) = \phi t^{\phi-1}$ (we assume Weibull distribution).
- In Scenario I, $C \sim N(6.5, 1)$. In Scenario II, all event times greater than 6 were censored.
- We have simulated training and testing data, both contain 300 subjects.
- We have assessed the predictive performance (with IBS as well as EPCE) in the time interval $(t, t + \Delta] = (4, 5.5]$.

Results Scenario I

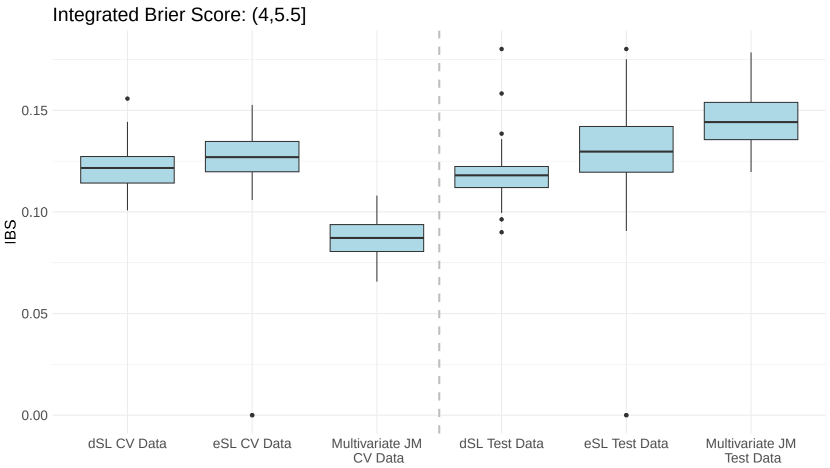


Figure: 100 datasets have been used. Results based on 3-fold CV.

Results Scenario I

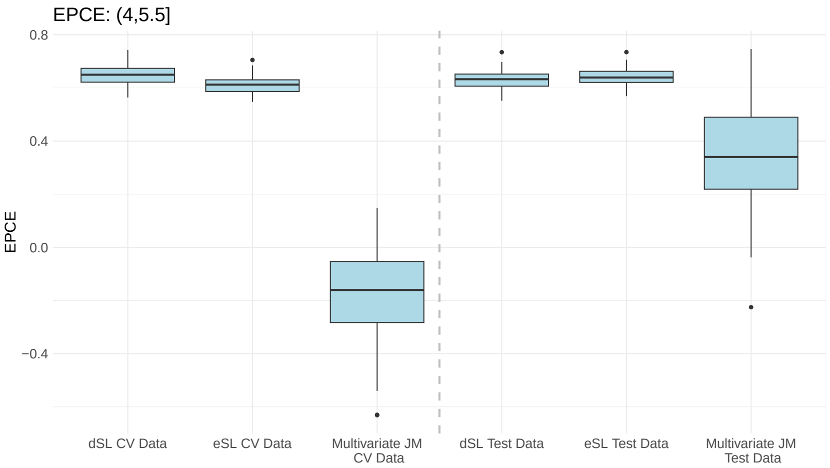


Figure: 100 datasets have been used. Results based on 3-fold CV.

Results Scenario II

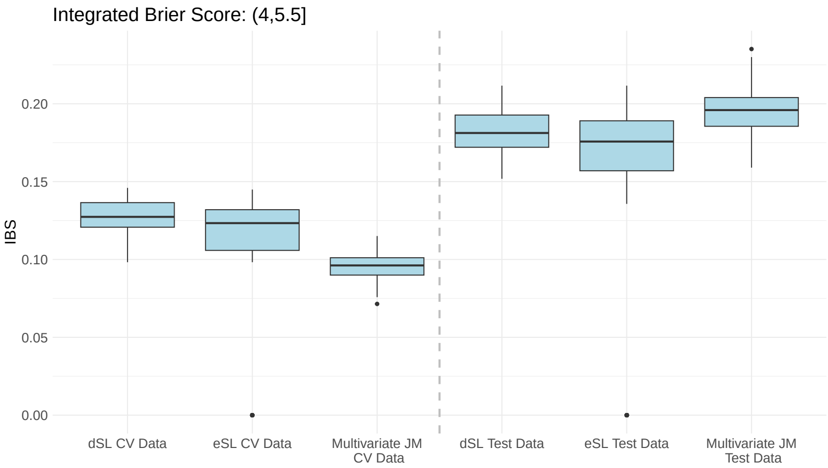


Figure: 100 datasets have been used. Results based on 3-fold CV.

Results Scenario II

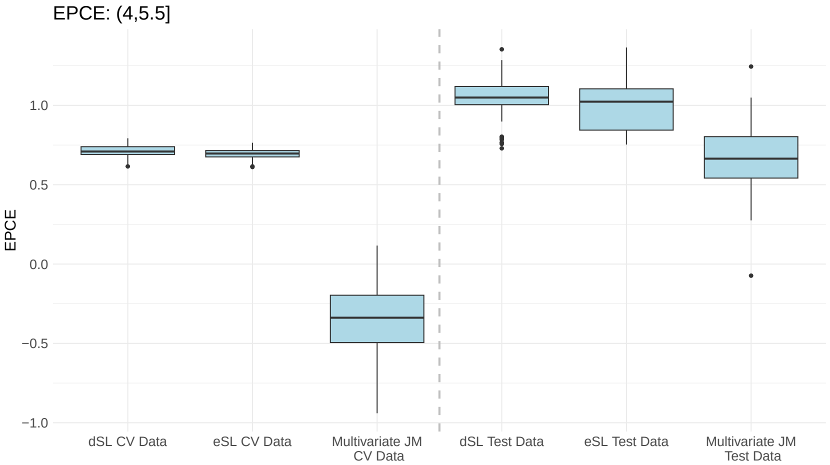


Figure: 100 datasets have been used. Results based on 3-fold CV.

Results

- In Scenario I, 38.32% and 36.81% of censoring has been obtained on average for training and testing data respectively.
- The discrete super learner has been Model 1 (first univariate joint model) in all the cases for the EPCE, and 95% for IBS.
- In Scenario II, on average, 42.05% and 41.58% of censoring for the training and testing data, respectively.
- In addition, the discrete super learner has been Model 2 (second univariate model) in all the 100 data sets for IBS, and in 93% of the data sets for EPCE.

Over-fitting analysis

We have calculated, for each of the simulated data sets, the difference between the proper scoring rule obtained by analyzing the testing data and the training data. We show the mean of those differences:

	Scenario 1			Scenario II		
	dSL	eSL	Multi JM	dSL	eSL	Multi JM
IBS	-0.0037	0.0041	0.0581	0.0547	0.0425	0.0993
EPCE	-0.0242	0.0306	0.4994	0.3242	0.2834	0.94521

Differences between SL and multiJM

We have calculated for each simulated data set the difference between the accuracy of the dSL and the eSL and that of the multivariate joint model.

- The average of these differences for Scenario I:

	Train		Test	
	$dSL - mJM$	$eSL - mJM$	$dSL - mJM$	$eSL - mJM$
IBS	0.0345	0.0349	-0.0272	-0.0191
EPCE	0.8015	0.7627	0.2779	0.2938

Differences between SL and multiJM

- The average of these differences for Scenario II:

	Train		Test	
	$dSL - mJM$	$eSL - mJM$	$dSL - mJM$	$eSL - mJM$
IBS	0.0316	0.0031	-0.0130	-0.0536
EPCE	1.0004	0.9772	0.3795	0.3153

Comments

- In general, the IBS results have been more stable than those of EPCE. In fact, in several data sets (32 and 55, in first and second scenario, respectively) when calculating the EPCE for the multivariate joint model on the test data we have obtained an `Inf`.
- There is much more variability in the EPCE of the multivariate joint model.

FPCA-based methods

- Functional data analysis provides a statistical framework that is flexible and well suited to model sparsely sampled longitudinal data, and time-to-event outcomes.
- To deal with our problem we will use multivariate functional principal component analysis (MFPCA).

Restrictions

- We assume that the response vector $\boldsymbol{y}_i$ conditional on the random effects $\boldsymbol{b}_i$ has a continuous distribution $\mathcal{F}_{\Psi}$.
- Although a normality assumption is not required, the longitudinal outcomes must be continuous.

Assumptions

- We assume that the observed data $\boldsymbol{y}_{li}$ is a noisy measurement of the latent outcome process $X_{li}(t)$, where time $t \in \mathcal{T} = [0, \tau]$ and $\tau = max\{T_i^* : i = 1, \ldots, n\}$.
- This is, we end up with $y_{li}(t_{ij}) = X_{li}(t_{ij}) + \varepsilon_{li}(t_{ij})$, where $\varepsilon_{li}(t_{ij})$ are independent errors centered and with variance $\sigma^2_{\varepsilon_l}$.

Scores computation

A two-step procedure is used to compute the MFPC scores.

1. We can approximate the latent process using the Karhunen Loève expansion:

$$X_{li}(t) = \mu_l(t) + \sum_{r=1}^{\infty} \xi_{ilr}\phi_{lr}(t),$$

Where $\mu_l(t)$ is the unknown smoothed mean function of $X_{li}(t)$. $\{\xi_{ilr}\}_{r=1,\ldots,\infty}$ are the so-called FPC scores. They are uncorrelated random variables with mean zero and variance λ_{lr}. $\{\phi_{lr}(t)\}_{r=1,\ldots,\infty}$ are orthonormal eigenfunctions. The last two are coming from the spectral decomposition of the covariance function of $X_{li}(t)$ (Mercer's Theorem).

Scores computation

1. We apply the PACE algorithm to all L longitudinal outcomes and estimate the eigenfunctions $\hat{\phi}(t) = (\hat{\phi}_{l1}(t), \dots, \hat{\phi}_{lR_l}(t))$ and the FPC scores $\hat{\boldsymbol{\xi}}_{li} = (\hat{\xi}_{il1}, \dots, \hat{\xi}_{ilR_l})$, where R_l is a proper truncation for the *l*th longitudinal outcome (chosen using percentage of variance explained (PVE)).

   We denote the vector of estimated FPC scores across all longitudinal outcomes for the *i*th subject as $\hat{\boldsymbol{\xi}}_i = (\hat{\boldsymbol{\xi}}_{1i}, \dots, \hat{\boldsymbol{\xi}}_{li})$, the length of the previous vector is denoted as $R_+ = \sum_{l=1}^{L} R_l$.

Scores computation

2. MFPCA implicitly models the correlations among outcomes by means of the correlations among the FPC scores. Estimates for the multivariate eigenfunctions for the *l*th outcome are given by

$$\hat{\Psi}_{lk}(t) = \sum_{r=1}^{R_l} [\hat{\boldsymbol{c}}_k]_r^{(l)} \hat{\phi}_{lr}(t),$$

Estimates for the MFPC scores of subject *i* can be computed as

$$\hat{\rho}_{ik} = \sum_{l=1}^{L} \sum_{r=1}^{R_l} [\hat{\boldsymbol{c}}_k]_r^{(l)} \hat{\xi}_{ilr}.$$

Scores computation

2. Lastly, the lth longitudinal outcome, $X_{li}(t)$, can be approximated by selecting the first $R^* \leq R_+$ scores and eigenfunctions based on PVE, and computing

$$E(y_{li}(t)) = \hat{X}_{li}(t) \approx \hat{\mu}_l(t) + \sum_{k=1}^{R^*} \hat{\rho}_{ik} \hat{\Psi}_{lk}(t).$$

Implementation of score calculation

- Implementation of MFPCA can be found in the R package `MFPCA`.
- The current implementation requires the longitudinal data to lie on a grid with fixed intervals.
- The observed longitudinal data $y_{li}(t_{ij})$ is first rounded to the nearest time corresponding to the fixed grid.
- A $n \times J \times L$ input matrix will be used to deal with the longitudinal data, where J is the grid length needed to accommodate the largest observation time.

MFPCA-Cox model

MFPC scores $\hat{\boldsymbol{\rho}}_i$, can be used as predictors. Cox proportional hazards model will be used to model the hazard function of ith subject,

$$h_i\left(t \mid \mathcal{Y}_i(t), \boldsymbol{w}_i\right) = h_0(t) \exp\left\{\boldsymbol{\gamma}^\top \boldsymbol{w}_i + \boldsymbol{\beta}^\top \hat{\boldsymbol{\rho}}_i\right\},$$

where $h_0(t)$ is an unspecified baseline hazard function and $\boldsymbol{w}_i$ is a vector of baseline covariates with corresponding regression coefficients $\boldsymbol{\gamma}$. $\boldsymbol{\beta}$ is the vector of regression coefficients for the multivariate longitudinal predictors through the R^* estimated MFPC scores $\hat{\boldsymbol{\rho}}_i$.

MFPCA-Cox model

To predict the risk of an event not occurring within a time window $(t, u]$, we use the conditional probability of event-free at any time u after t

$$\begin{aligned}\pi_j(u \mid t) &= P\left(T_j^* \geq u \mid T_j^* > t, \mathcal{Y}_j(t), \gamma, \boldsymbol{\beta}, \hat{\boldsymbol{\rho}}_i\right) \\ &= \left(\frac{\hat{S}_0(u)}{\hat{S}_0(t)}\right)^{\exp\{\hat{\boldsymbol{\gamma}}^\top \boldsymbol{w}_i + \hat{\boldsymbol{\beta}}^\top \hat{\boldsymbol{\rho}}_i\}},\end{aligned}$$

where $\hat{S}_0(t) = \exp\{-\int_0^s \hat{h}_0(t)dt\}$ is the baseline survival function. $\hat{h}_0(t)$ is estimated via Breslow estimator.

MFPCA-DeepSurv model

- DeepSurv is a feed-forward neural network which predicts the effects of a patient's covariates on their hazard rate parameterized by the weights of the network $\boldsymbol{\theta}$.
- DeepSurv is able to learn complex and nonlinear relationships between patient's covariates and their event risk.
- The output of the network $\hat{g}_{\boldsymbol{\theta}}(w)$ is a single node with a linear activation which estimates the log-hazard function in the following Cox model

$$h(t|x) = h_0(t)\exp\{g(w)\}.$$

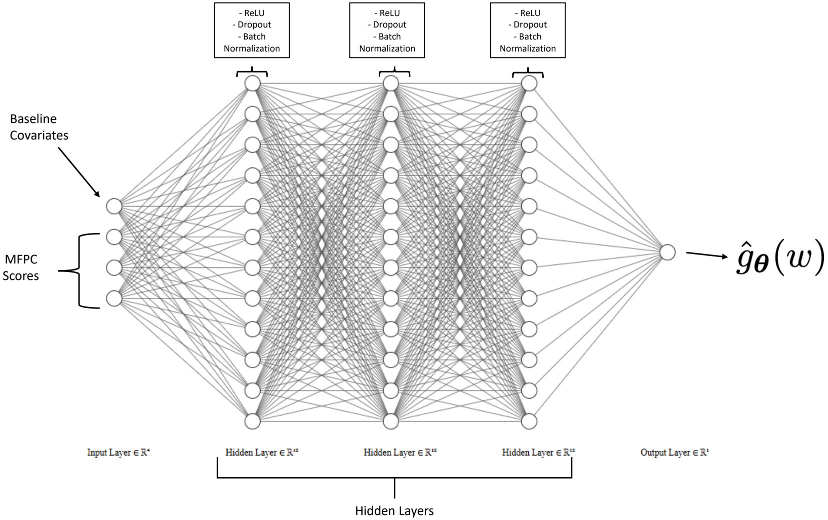
MFPCA-DeepSurv model

- DeepSurv is trained on a loss function based on the Cox partial likelihood function. In particular, the model minimizes the average negative log partial likelihood

$$l(\theta) = -\frac{1}{N_{\delta=1}} \sum_{i|\delta_i=1} \left(\hat{g}_{\boldsymbol{\theta}}(w_i) - \log \left(\sum_{j \in \mathcal{R}(T_i^*)} \exp(\hat{g}_{\boldsymbol{\theta}}(w_j)) \right) \right),$$

  where $N_{\delta=1}$ is the total number of subjects who had suffered the event and $\mathcal{R}(t)$ is the set of patients at risk at time t. Gradient descent optimization is used to find the weights of the network minimizing it.

MFPCA-DeepSurv model


- ReLU
- Dropout
- Batch Normalization
- ReLU
- Dropout
- Batch Normalization
- ReLU
- Dropout
- Batch Normalization
Baseline Covariates
MFPC Scores
$\hat{g}_{\boldsymbol{\theta}}(w)$
Input Layer $\in \mathbb{R}^4$
Hidden Layer $\in \mathbb{R}^{12}$
Hidden Layer $\in \mathbb{R}^{12}$
Hidden Layer $\in \mathbb{R}^{12}$
Output Layer $\in \mathbb{R}^1$
Hidden Layers

MFPCA-DeepSurv model

- Each linear layer in the feed-forward neural network is followed by an activation function. Rectified linear unit (ReLU) is the chosen. Where $ReLU(x) = max(0, x)$.
- The latter is followed by the dropout layer. Used to prevent our feed-forward neural network from over-fitting.
- Finally, batch normalization is used in order to re-center and re-scale the layer output to allow for faster and more stable training.
- Adaptive moment estimation (Adam) is used for the gradient descent algorithm.

MFPCA-DeepSurv model

- Furthermore, once we have the output of the feed-forward neural network $\hat{g}_{\boldsymbol{\theta}}(w)$, we estimate $\hat{h}_0(t)$ via Breslow estimator.
- Thus, individualized dynamic predictions can be computed using

$$\begin{aligned}\pi_j(u \mid t) &= P\left(T_j^* \geq u \mid T_j^* > t, \boldsymbol{\theta}\right) \\ &= \left(\frac{\hat{S}_0(u)}{\hat{S}_0(t)}\right)^{\exp\{\hat{g}_{\boldsymbol{\theta}}(w)\}} .\end{aligned}$$

Simulation study

The multivariate joint model is the only data-generating model. We consider $L = 5$ continuous longitudinal processes.

- Scenario I: Random censoring.
- Scenario II: Type I (administrative) censoring.

Longitudinal processes

For the longitudinal outcomes, we consider $L = 5$ longitudinal processes. All of them are generated via LLMs:

$$
\begin{aligned}
y_{1i}(t_{ij}) &= m_{1i}(t_{ij}) + \varepsilon_{1i}(t_{ij}) \\
&= (\beta_0^1 + b_{0i}^1) + (\beta_1^1 + b_{1i})t_{ij} + \beta_2^1 \texttt{sex}_i + \beta_3^1 \texttt{sex}_i t_{ij} + \varepsilon_{1i}(t_{ij}), \\
y_{2i}(t_{ij}) &= m_{2i}(t_{ij}) + \varepsilon_{2i}(t_{ij}) \\
&= (\beta_0^2 + b_{0i}^2) + (\beta_1^2 + b_{1i})t_{ij} + \beta_2^2 \texttt{treatment}_i + \beta_3^2 \texttt{treatment}_i t_{ij} + \varepsilon_{2i}(t_{ij}), \\
y_{3i}(t_{ij}) &= m_{i3}(t_{ij}) + \varepsilon_{3i}(t_{ij}) \\
&= (\beta_0^3 + b_{0i}^3) + (\beta_1^3 + b_{1i}^3)t_{ij} + \varepsilon_{3i}(t_{ij}), \\
y_{4i}(t_{ij}) &= m_{4i}(t_{ij}) + \varepsilon_{4i}(t_{ij}) \\
&= (\beta_0^4 + b_{0i}^4) + (\beta_1^4 + b_{1i}^4)t_{ij} + \beta_2^4 \texttt{sex}_i + \varepsilon_{4i}(t_{ij}), \\
y_{5i}(t_{ij}) &= m_{5i}(t_{ij}) + \varepsilon_{5i}(t_{ij}) \\
&= (\beta_0^5 + b_{0i}^5) + (\beta_1^5 + b_{1i}^5)t_{ij} + \beta_2^5 \texttt{treatment}_i + \varepsilon_{5i}(t_{ij}).
\end{aligned}
$$

Survival outcomes

The multivariate joint model:

$$h_i\left(t \mid \mathcal{Y}_i(t), \boldsymbol{w}_i\right) = h_0(t)\exp\left\{\gamma_0 + \gamma_1 \mathtt{sex}_i + \sum_{l=1}^{5} \alpha_l m_{li}(t)\right\},$$

The five univariate joint models that make up the library $\mathcal{L} = \{M_1, M_2, M_3, M_4, M_5\}$:

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

Some model specifications

- Event times have been simulated by means of the inverse transform sampling method.
- The baseline hazard function is $h_0(t) = \phi t^{\phi-1}$ (we assume Weibull distribution).
- In Scenario I, $C \sim exp$ with rate $1/10$. In Scenario II, all event times greater than 5 were censored.
- We have simulated training and testing data, both contain 300 subjects.
- We have assessed the predictive performance using the IBS with IPCW in the time interval $(t, t + \Delta] = (2.5, 3.5]$.

Some model specifications

A constraint has been added to use MFPCA-based models:

- From each model and for each subject we simulated longitudinal responses at time zero and then at 16 randomly selected time points coming from $U(0, 10)$.
- In addition, a constraint has been used when selecting the time points, a minimum distance of 0.5 between randomly selected points is required.
- The latter is because when using the grid of fixed time intervals to compute MFPC scores, a grid with a distance of 0.5 between time points is used.

Model specification for the MFPCA-based models

- To compute the MFPCA, a grid with fixed time intervals has been used. These time intervals occurs every 0.5 units of time.
- An 80% of PVE has been used. The latter is used to compute the number of principal components to be computed.
- In both MFPCA-based models, Breslow estimator has been used to estimate the baseline hazard function.

For the MFPCA-DeepSurv model, the hyperparameters of the feed-forward neural network used are:

- 2 hidden layers, 64 nodes per hidden layer.
- 0.2 dropout probability.
- 12 number of epochs.
- Batch size of 32.

Results Scenario I

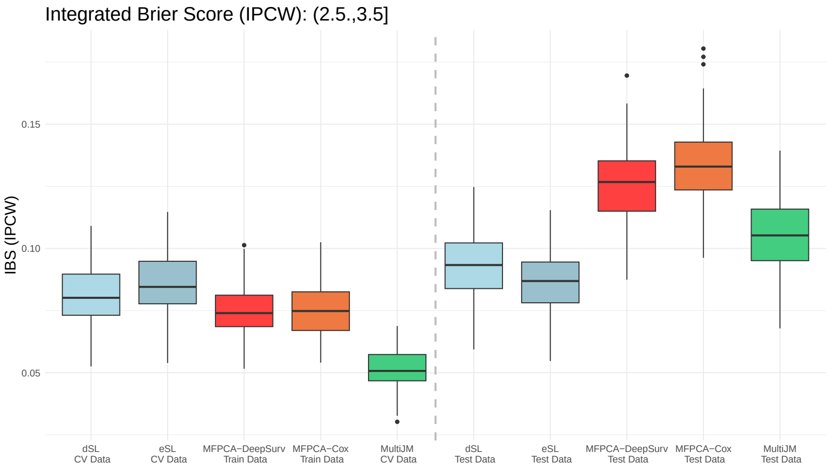


Figure: 100 datasets have been used. Results related with SL are based on 3-fold CV.

Results Scenario II

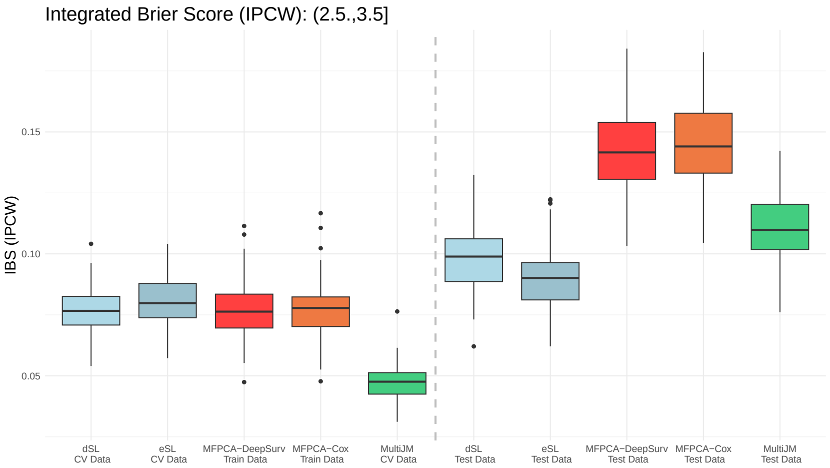


Figure: 100 datasets have been used. Results related with SL are based on 3-fold CV.

Results

- In Scenario I, on average, we have obtained 41.12% and 41.94% of censoring in training and testing data, respectively.
- In 93% of simulated datasets the discrete super learner is the univariate Model 2, in the remaining data sets is either the univariate Model 5 (6%) or Model 1 (1%).
- In Scenario II, on average 40.48% and 40.82% of censoring have been obtained in training and testing data, respectively.
- In the 94% of data sets the dSL is the univariate joint Model 2, in the rest of cases the univariate Model 5.

Results

- Predictions based on the super learning algorithm seem to better account for over-fitting.
- In both scenarios eSL is giving the most accurate prediction in testing data.
- An advantage of the MFPCA-based models with respect to super learning procedures is the computation time.

BMA: indicator approach

- A widely used approach in Bayesian analysis to do variable selection, is to introduce an indicator random variable I_i for each covariate, and introduce these into the model in order to ’zero out’ inactive covariates. The latter can be used also with BMA.
- In particular, in our case we can rewrite the the multivariate joint model as

$$h_i\left(t \mid \mathcal{Y}_i(t), \boldsymbol{w}_i\right) = h_0(t) \exp\left\{\gamma^\top \boldsymbol{w}_i + \sum_{l=1}^{L} I_l f_l(t, \mathcal{Y}_{li}(t), \boldsymbol{b}_{li}, \alpha_l)\right\},$$

where $I_l, l = 1, \ldots, L$ are indicator random variables, with prior distribution $Bernoulli(p)$, where $p \sim Beta(a, b)$, a, b known parameters.

BMA: indicator approach

- Indicators approach is not feasible to do BMA in our case, since we want to avoid fitting the multivariate joint model. With this approach we not only adjust the multivariate model, but we also add L parameters to be taken into account.
- The approach above can be useful to do a comparison with the super learning algorithm with small L. We have to take into account that we are taking into account 2^L models (all the possible combinations of models with the L different longitudinal sub-models). Thus, one could expect that will be the model making the best predictions.
- I am not sure how having the association parameter α_l, can affect the new indicator random variables.