Super Learning in Joint Models

Arnau G. Fernández

MESIO UPC-UB

Director: Dimitris Rizopoulos

Co-Director: Guadalupe Gómez

Final Master's Thesis

Contents

Problem and objective

- Problem: Fitting multivariate joint models is computationally expensive. When having a large number of longitudinal biomarkers, can we use Super Learning to compute predictions?
- Objective: The main objective of this work is to demonstrate that the predictions calculated via Super Learning are good enough to be used instead of the predictions of the true model (the multivariate joint model).

Problem and objectives

To achieve the objective, we use three different approaches:

- Theoretical approach: we want to demonstrate, from a theoretical point of view, that the information provided by different univariate joint models approximate, in some manner, the information provided by a multivariate joint model.
- Case study: we will use real data to demonstrate that with Super Learning procedures we obtain a predictive performance similar to that obtained with the multivariate model.
- Simulation study: a simulation study has been carried out to demonstrate the above.

Introduction: assumptions and notation

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 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, $\boldsymbol{y}_{il}$ is the $n_{il} \times 1$ 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:

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

where:

- $g_l(\cdot)$ denotes a known one-to-one monotonic link function for the lth longitudinal outcome
- $\boldsymbol{x}_{il}(t)$ and $\boldsymbol{z}_{il}(t)$ denote the time-dependent design vectors for the fixed-effects β and for the random effects $\boldsymbol{b}_{il}$
- ϕ denote the scale parameter
- Random effects follow a multivariate normal with mean zero and variance-covariance matrix D_l

Introduction: assumptions and notation

For the survival process, we have

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

where:

- $\mathcal{Y}_{il}(t) = \{m_{il}(s), 0 \leq s < t\}$ denotes the history of the underlying lth longitudinal process up to t
- $h_0(\cdot)$ 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 $v_1, \ldots, v_Q$,
- γ_{h_0} the vector of spline coefficients.

Introduction: assumptions and notation

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

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

where:

- $vech(\boldsymbol{D}_1), \ldots, vech(\boldsymbol{D}_L)$ denotes the unique elements of the variance-covariance matrices $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_L$.
- Normal priors for all the regression coefficients, and inverse-gamma priors for $\phi_1, \ldots, \phi_L$ and the diagonal elements of $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_L$, and LKJ prior for the correlation matrices of the random effects.

Introduction: assumptions and notation

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:

$$M1: \quad h_i(t \mid \mathcal{Y}_{i1}(t), \boldsymbol{w}_i) = h_0(t) \exp\left\{\boldsymbol{\gamma}^\top \boldsymbol{w}_i + \alpha_1 f_1(t, \mathcal{Y}_{i1}(t), \boldsymbol{b}_{i1})\right\},$$

$$M2: \quad h_i(t \mid \mathcal{Y}_{i2}(t), \boldsymbol{w}_i) = h_0(t) \exp\left\{\boldsymbol{\gamma}^\top \boldsymbol{w}_i + \alpha_2 f_2(t, \mathcal{Y}_{i2}(t), \boldsymbol{b}_{i2})\right\},$$

$$\ldots$$

$$M_L: \quad h_i(t \mid \mathcal{Y}_{iL}(t), \boldsymbol{w}_i) = h_0(t) \exp\left\{\boldsymbol{\gamma}^\top \boldsymbol{w}_i + \alpha_L f_L(t, \mathcal{Y}_{iL}(t), \boldsymbol{b}_{iL})\right\}.$$

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

Theoretical approach: Assumptions

Let us assume prior independence:

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

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

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

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 $f_l(t, \mathcal{Y}_{il}(t), \boldsymbol{b}_{il}) = m_{il}(t)$ for all $l = 1, \ldots, L$.

Likelihood of the multivariate JM

$$p(\boldsymbol{y}_i, T_i, \delta_i \mid \boldsymbol{b}_i, \boldsymbol{\theta}) = p(\boldsymbol{y}_i \mid \boldsymbol{b}_i, \boldsymbol{\theta})\, p(T_i, \delta_i \mid \boldsymbol{b}_i, \boldsymbol{\theta})$$

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

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

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

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

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

Likelihood of the product of the univariate JM's

$$\prod_{l=1}^{L} p\left(\boldsymbol{y}_{il}, T_i, \delta_i \mid \boldsymbol{\theta}, \boldsymbol{b}_i\right)$$

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

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

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

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

Main differences between the likelihoods

- The longitudinal contribution is identical.
- The random effects contribution is identical.
- The survival outcome contribution is different:
  - L factors arise.
  - $\sum_{l=1}^{L} \exp\left(\alpha_l m_{il}(s)\right)$, the sum out of the exponential.

Main differences between the priors

Multivariate joint model prior:

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

Product of univariate joint model priors:

$$\begin{aligned} p(\boldsymbol{\theta}) =& p(\boldsymbol{\beta}_1)\cdots p(\boldsymbol{\beta}_L)p(\boldsymbol{\phi}_1)\cdots p(\boldsymbol{\phi}_L)p(\boldsymbol{\alpha}_1)\cdots p(\boldsymbol{\alpha}_L) \\ & \times p(vech(\boldsymbol{D}_1))\cdots p(vech(\boldsymbol{D}_L))p(\tau)^L p(\boldsymbol{\gamma}_{h_0})^L p(\boldsymbol{\gamma})^L. \end{aligned}$$

Main differences between the priors

For those priors related with the

- longitudinal process, the contribution is identical.
- random effects, the contribution is identical.
- survival outcome, the contribution changes.

What's next?

- What about the independence/correlation between longitudinal processes?
- The fact that in the *L* models we use the same specification for the survival process implies that we are somehow collecting more information about the survival process as L grows?
- How do the differences encountered affect?

Mayo Clinic Primary Biliary Cirrhosis Data analysis

The *Mayo Clinic Primary Biliary Cirrhosis Data* (`pbc2`) is a data set available in 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. Goals of the case study:

- Become familiar with the functions of the *JMbayes2* package.
- See with real data that SL allows us to calculate predictions with an accuracy similar to that of the multivariate joint model.

Longitudinal outcomes

We consider $L = 5$ longitudinal outcomes:

$$\begin{aligned}\log(\texttt{serBilir}(t_{ij})) &= m_{i1}(t_{ij}) + \varepsilon_{i1}(t_{ij}) \\ &= (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_{ij} + \varepsilon_{i1}(t_{ij}),\end{aligned}$$

$$\begin{aligned}\texttt{prothrombin}(t_{ij}) &= m_{i2}(t_{ij}) + \varepsilon_{i2}(t_{ij}) \\ &= (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_{ij} + \beta_2 \texttt{sex}_i + \beta_3 \texttt{sex}_i * t_{ij} + \varepsilon_{i2}(t_{ij}),\end{aligned}$$

$$\begin{aligned}\log\left(\frac{p(\texttt{ascites}(t_{ij}) = 1)}{1 - p(\texttt{ascites}(t_{ij}) = 1)}\right) &= m_{i3}(t_{ij}) + \varepsilon_{i3}(t_{ij}) \\ &= (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_{ij} + \beta_2 \texttt{sex}_i + \varepsilon_{i3}(t_{ij})\end{aligned}$$

Longitudinal outcomes

$$\begin{aligned}\texttt{albumin}(t_{ij}) &= m_{i4}(t_{ij}) + \varepsilon_{i4}(t_{ij}) \\ &= (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_{ij} + \varepsilon_{i4}(t_{ij}),\end{aligned}$$

$$\begin{aligned}\log\left(\frac{p(\texttt{hepatomegaly}(t_{ij}) = 1)}{1 - p(\texttt{hepatomegaly}(t_{ij}) = 1)}\right) &= m_{i5}(t_{ij}) + \varepsilon_{i5}(t_{ij}) \\ &= (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_{ij} + \varepsilon_{i5}(t_{ij}).\end{aligned}$$

Where time is in years. We have three linear mixed models (LMM) and two generalized linear mixed models (GLMM).

Multivariate and univariate JM

For the multivariate joint model we have:

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

We have the library $\mathcal{L} = \{M_1, M_2, M_3, M_4, M_5\}$ consisting of the following $L = 5$ univariate joint models:

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

IBS and EPCE

- We have used the integrated Brier score (IBS) and the expected predictive cross-entropy (EPCE) as predictive performance metrics.
- We evaluate the predictive performance in two time intervals $(t, t + \Delta t]$, say $(5, 8]$ and $(7, 10]$.
- $(5, 8]$: 202 individuals were at risk at year 5, besides 28 events, 70 censored times occurred in the interval $(5, 8]$.
- $(7, 10]$: 129 individuals were at risk at time 7, 22 events occurred in $(7, 10]$ and there were 56 censored times in that interval.

Results: IBS

	$(5, 8]$		$(7, 10]$	
	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: 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.

Results: EPCE

	$(5, 8]$		$(7, 10]$	
	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: 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.

Simulation study

We consider $L = 5$ longitudinal processes.

- Scenario I: one data-generating model, a multivariate joint model.
- Scenario II: six data-generating models, a multivariate joint model and $L = 5$ univariate joint models.

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:

$$
\begin{aligned}
y_{i1}(t_{ij}) &= m_{i1}(t_{ij}) + \varepsilon_{i1}(t_{ij}) \\
&= (\beta_0^1 + b_{0i}^1) + (\beta_1^1 + b_{1i})t_{ij} + \beta_2^1 \texttt{sex}_i + \beta_3^1 \texttt{sex}_i * t_{ij} + \varepsilon_{i1}(t_{ij}), \\
y_{i2}(t_{ij}) &= m_{i2}(t_{ij}) + \varepsilon_{i2}(t_{ij}) \\
&= (\beta_0^2 + b_{0i}^2) + (\beta_1^2 + b_{1i}^2)t_{ij} + \beta_2^2 \texttt{sex}_i + \varepsilon_{i2}(t_{ij}), \\
y_{i3}(t_{ij}) &= m_{i3}(t_{ij}) + \varepsilon_{i3}(t_{ij}) \\
&= (\beta_0^3 + b_{0i}^3) + (\beta_1^3 + b_{1i}^3)t_{ij} + \varepsilon_{i3}(t_{ij}),
\end{aligned}
$$

Longitudinal processes

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

Survival outcomes

The multivariate joint model:

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

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

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

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).
- The censored mechanism used was fixed Type I, all event times greater than 25 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 $(t, t + \Delta] = (12, 18]$.

Metrics computed

The predictive performance metrics (IBS and EPCE) have been computed for:

- The multivariate joint model, in the training and testing data.
- The ensemble Super Learner (eSL), in the training and testing data.
- The discrete Super Learner (dSL), in the training data.

Results Scenario I

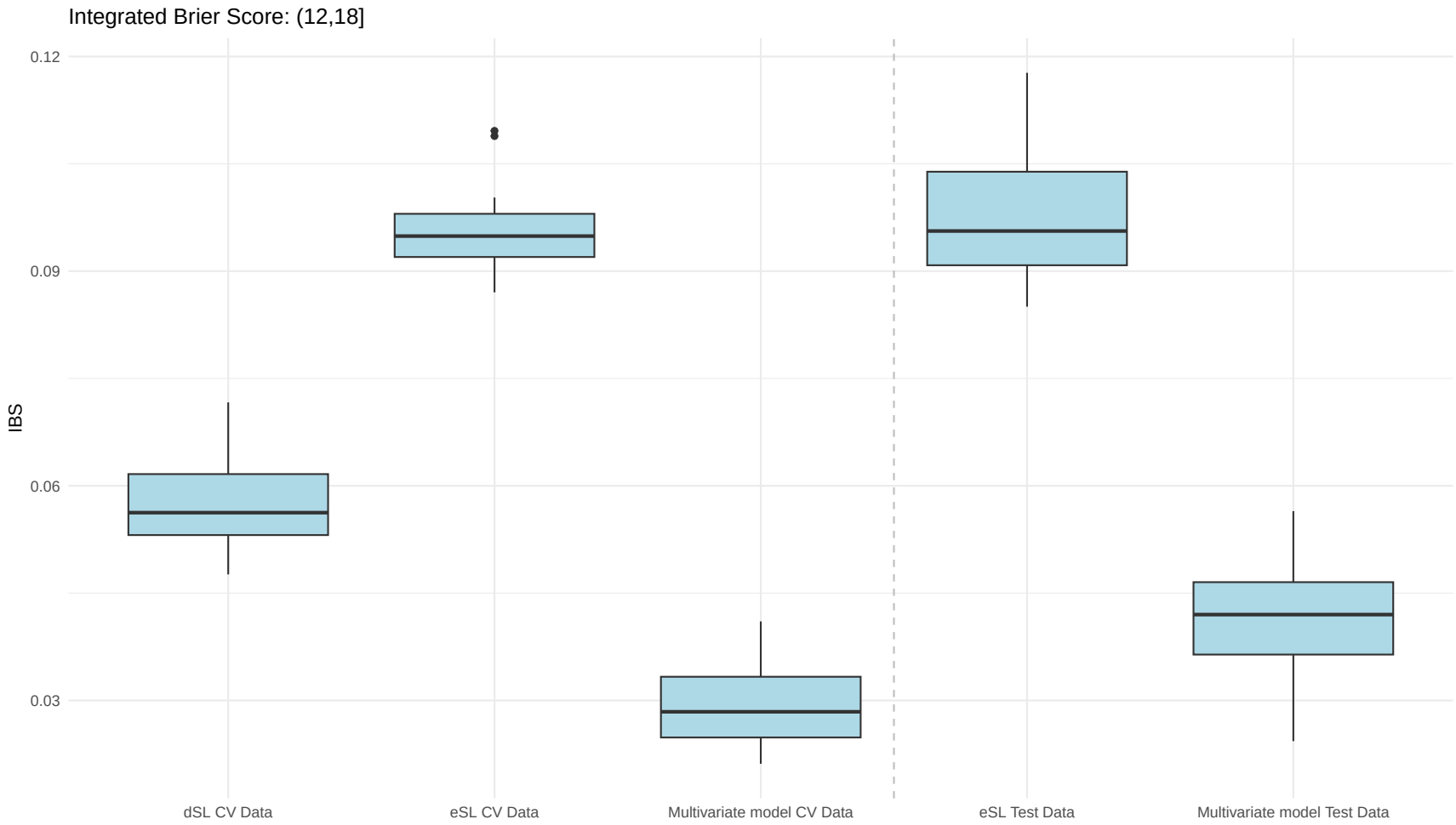

Figure: 20 datasets have been used. Results based on 3-fold CV.

Results Scenario I

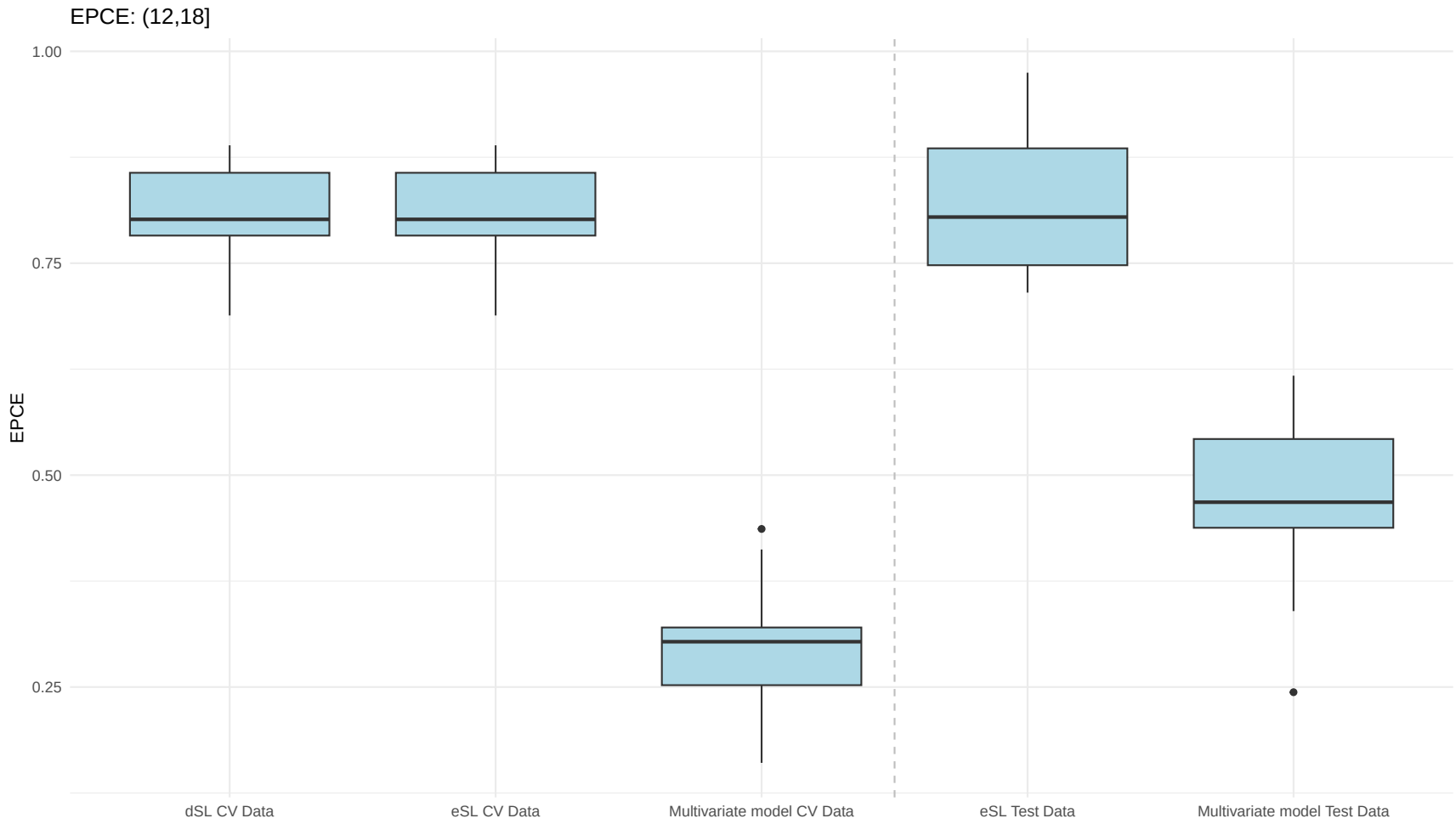

Figure: 20 datasets have been used. Results based on 3-fold CV.

Results Scenario II

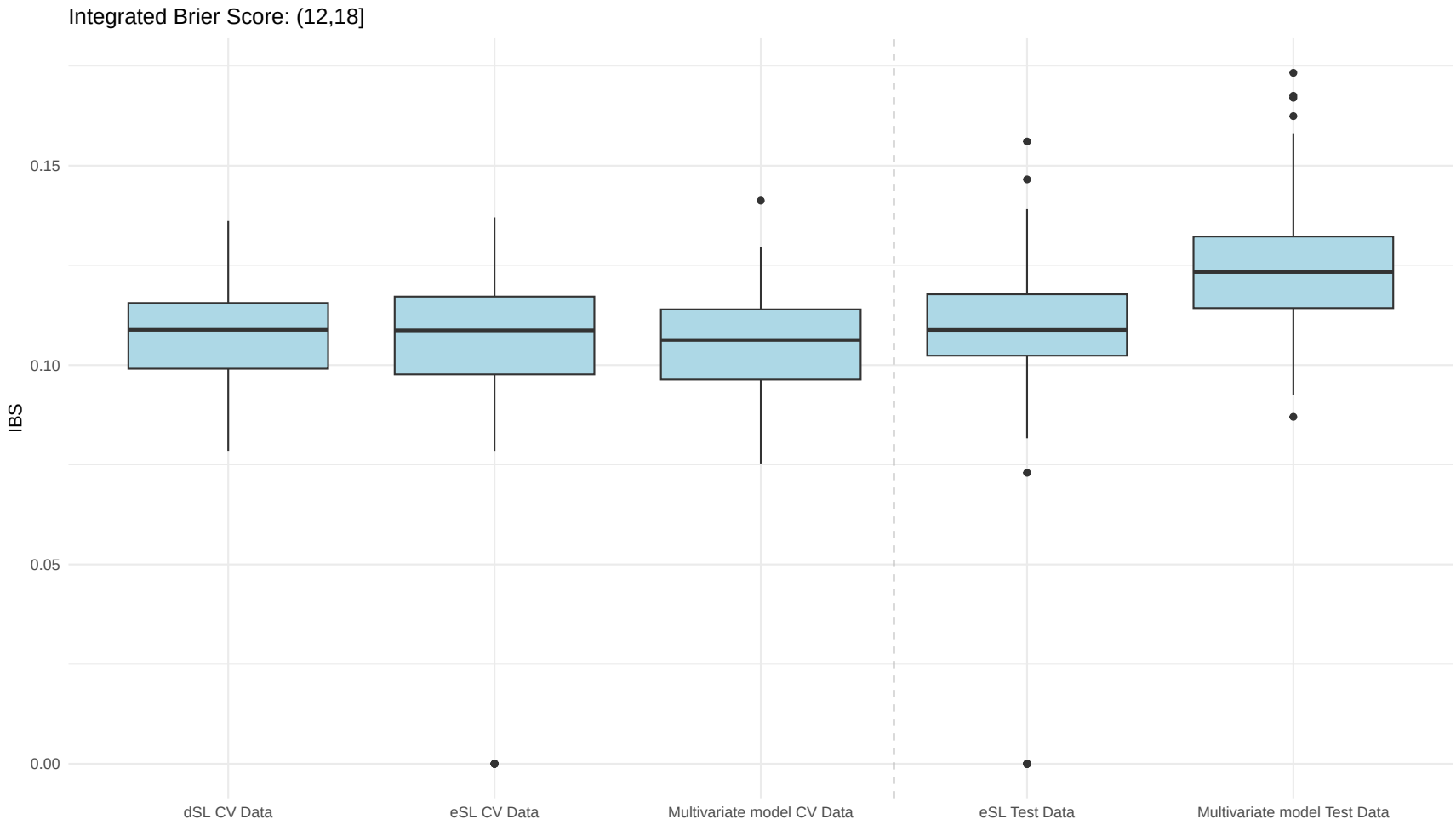

Figure: 100 datasets have been used. Results based on 3-fold CV.

Results Scenario II

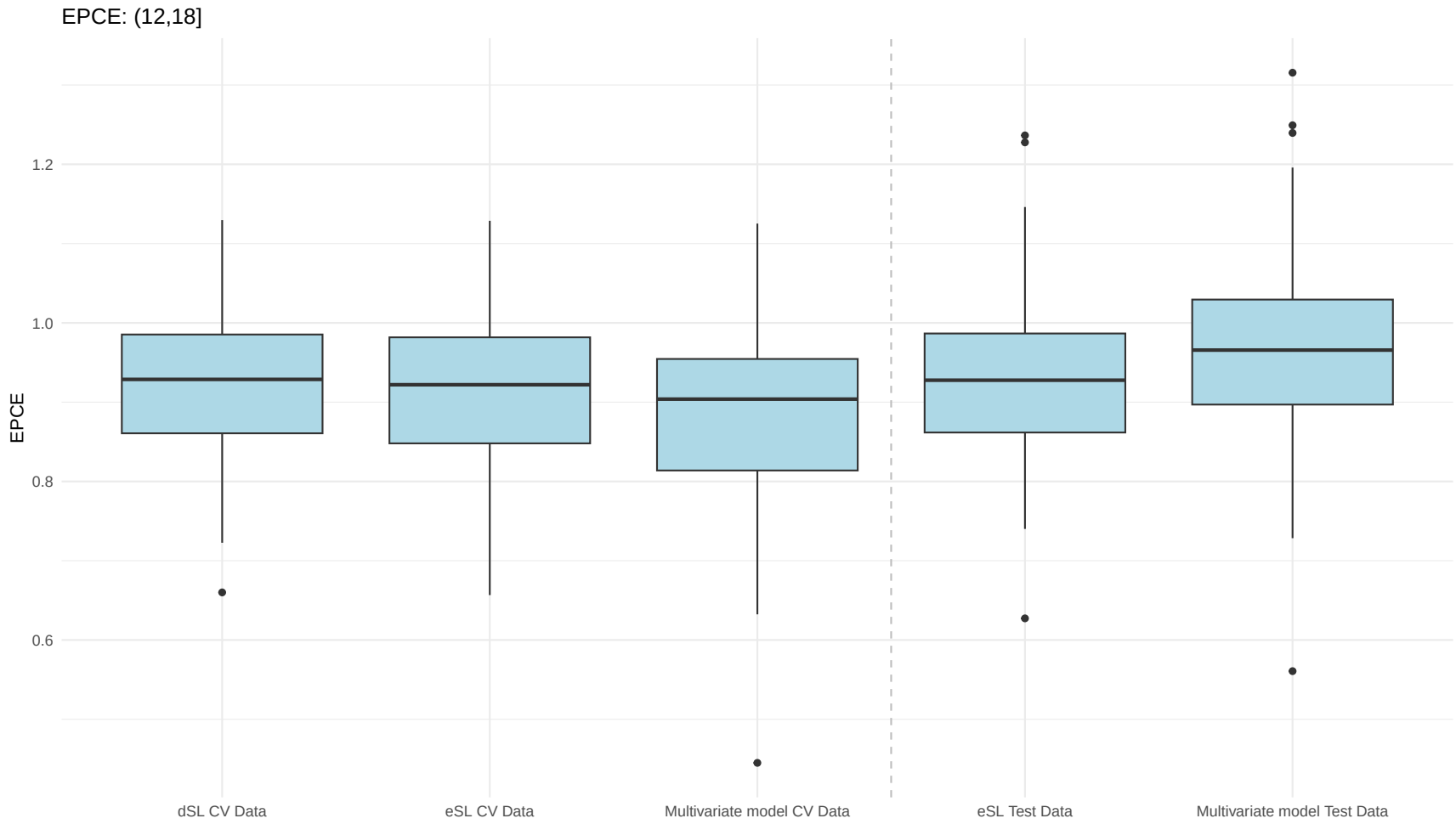

Figure: 100 datasets have been used. Results based on 3-fold CV.

Simulation study conclusions

- It seems that the results obtained in Scenario I are not good enough. Could there be overfitting by using only the multivariate model?
- Results in Scenario II seems to be better.

Questions and comments: theoretical part

- I think that demonstrating that there is a relationship between the posterior distribution of the multivariate joint model and the posteriors of the univariate joint models, and specify as much as possible how this relationship is, can be a good theoretical approach to see that we can approximate multivariate models via univariate models.
- Analyzing what happens with posteriors within the SL algorithm may be much more complicated.

Questions and comments: case study

- In which database can we do the study? Is there a database in which you have a special motivation to do the case study?

Questions and comments: simulation study

- Is the % of censoring fixed in [1]?
- Which is the oracle model exposed in the Simulation Results in [1]?
- How can I compute dSL in test data?
- What other scenarios might be interesting to explore?

References


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