Recent Advances in Joint Models for Multivariate Longitudinal Data and Event-times with Application to Cancer

Ruwanthi Kolamunnage-Dona

Department of Biostatistics, Institute of Translational Medicine

kdrr@liverpool.ac.uk

24-25th January 2019, ISPED, Bordeaux, France

Motivation for the extended joint model

- In health research, often a relatively large number of quantities are measured over patients' follow-up over time in order to fully explore the damage caused by adverse clinical events
- Harnessing all available information in a single model leads to improved estimation and prediction.

Data for the extended joint model

For each individual $i = 1, \ldots, n$, we observe

- $y_i = (y_{i1}^\top, \ldots, y_{iK}^\top)$ is the K-variate continuous outcome vector, where each y_{ik} denotes an $(n_{ik} \times 1)$-vector of observed longitudinal measurements for the k-th outcome type: $y_{ik} = (y_{i1k}, \ldots, y_{in_{ik}k})^\top$
- Observation times t_{ijk} for $j = 1, \ldots, n_{ik}$, which can differ between individuals and outcomes
- (T_i, δ_i), where $T_i = \min(T_i^*, C_i)$, where T_i^* is the true event time, C_i corresponds to a potential right-censoring time, and δ_i is the failure indicator equal to 1 if the failure is observed $(T_i^* \leq C_i)$ and 0 otherwise.

Longitudinal data sub-model

A multivariate or K-variate process, and for the k-th outcome $(k = 1, \ldots, K)$

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

where

- $\mu_{ik}(t) = X_{ik}^{\top}(t)\beta_k$ is the mean response
- $X_{ik}(t)$ is a p_k-vector of covariates (possibly time-varying) with corresponding fixed effect terms β_k
- $W_{1i}^{(k)}(t)$ is a zero-mean *latent* Gaussian process
- $\varepsilon_{ik}(t)$ is the model error term, which is i.i.d. $N(0, \sigma_k^2)$ and independent of $W_{1i}^{(k)}(t)$.

Time-to-event sub-model

Cox proportional hazards model,

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

where

- $\lambda_0(\cdot)$ is an unspecified baseline hazard function
- $V_i(t)$ is a q-vector of covariates with corresponding fixed effect terms γ_v
- $W_{2i}(t)$ is a zero-mean *latent* Gaussian process, independent of the censoring process.

Association structure

Defined by the link between $W_1^{(k)}(t)$ and $W_2(t)$; each $W_1^{(k)}(t)$ is a linear combination of random effects:

$$W_{1i}^{(k)}(t) = Z_{ik}^{\top}(t) b_{ik} \text{ where } \boldsymbol{b_i} \sim N(\mathbf{0}, \boldsymbol{D})$$

with

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

Model also captures

1. within-individual correlation between longitudinal measurements via $\text{var}(b_{ik}) = D_{kk}$
2. dependence between the different longitudinal outcomes via $\text{cov}(b_{ik}, b_{il}) = D_{kl}$ for $k \neq l$

Joint likelihood

The *observed* data likelihood is given by

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

where $\theta = (\beta^\top, \text{vech}(D), \sigma_1^2, \ldots, \sigma_K^2, \lambda_0(t), \gamma_v^\top, \boldsymbol{\gamma}_{\mathbf{y}}^\top)$ is the collection of unknown parameters that we want to estimate.

This can be calculated by rewriting

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

where $f(y_i \mid \theta) \sim N(X_i\beta, \Sigma_i + Z_i D Z_i^\top)$.

Estimation

We determine maximum likelihood estimates of θ using

- MCEM algorithm=EM algorithm + Monte Carlo (MC) E-step[1]
- Same as the conventional Expectation-Maximisation (EM) algorithm, except that
- E-step exploits a MC integration (instead of a Gaussian quadrature method) which is beneficial when the dimension of random effects becomes large

Starting values: use estimates from separate analyses of the longitudinal and event-time components.

[1] See Wei and Tanner (1990)

Monte Carlo E-step

- E-step calculates several multi-dimensional expectations of function of random effects

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

- Use Monte Carlo sampling to estimate the integrals and approximate the expectation by

$$\approx \frac{\frac{1}{N} \sum_{d=1}^{N} h\left(b_i^{(d)}\right) f\left(T_i, \delta_i \mid b_i^{(d)}; \hat{\theta}\right)}{\frac{1}{N} \sum_{d=1}^{N} f\left(T_i, \delta_i \mid b_i^{(d)}; \hat{\theta}\right)}$$

where $b_i^{(1)}, b_i^{(2)}, \ldots, b_i^{(N)}$ are a random sample from $b_i \mid y_i, \theta$.

Convergence

In MCEM framework, there are 2 complications to account for

1. false convergence declared due to chance
   $\Rightarrow$ **Solution**: require convergence for 3 consecutive iterations
2. estimators swamped by Monte Carlo error, thus precluding convergence
   $\Rightarrow$ **Solution**: increase Monte Carlo size N as algorithm moves closer towards maximizer

See Hickey et al. (2018) for more detail on this algorithm, restrictions on convergence (stopping rules) & our simulation investigations.

Dynamic prediction

We calculate the conditional survival probability for a new individual at time $u > t$ given that the individual survived up to time t and provided a set of longitudinal outcome measurements $\boldsymbol{y_t}$ until time t:

$$P[T^* \geq u \,|\, T^* > t, \boldsymbol{y_t}; \hat{\theta}] = \mathbb{E}\left[\frac{S\left(u \,|\, b; \hat{\theta}\right)}{S\left(t \,|\, b; \hat{\theta}\right)}\right]$$

where $\hat{\theta}$ denotes the estimated joint model, and $S(. \,|\, b; \hat{\theta})$ is the survival function.

It can be calculated using estimators proposed by Rizopoulos (2011), based on either a first-order approximation or Monte Carlo simulation.

Software

- We can implement all of this in the R package `joineRML`[2]
- Fit the model using `joineRML::mjoint()`
- Calculates approximate SEs by default, but bootstrap SEs available via `joineRML::bootSE()`
- Built-in functions to get dynamic predictions
- `joineRML` package can also be used to fit classical joint models, but using MCEM rather than EM optimisation

[2] Hickey et al. (2018)

Predicting early recurrence of HCC

- Hepatocellular carcinoma (HCC) is the most common type of primary liver cancer in adults; it is the sixth most common cause of cancer worldwide
- Hepatic resection is a well-accepted therapy for HCC, but majority of patients subsequently develop tumour recurrence
- A better risk assessment is quite important
- Attention has been directed towards HCC-specific biomarkers to use in the early identification.

Aim: build a tool that predicts risk of HCC recurrence for individual patients

NB. biomarker transformations chosen according to Box-Cox transformations

Proposed joint model for HCC data

Trivariate longitudinal outcome sub-model

$$\begin{aligned}
y_1 = \log(\texttt{AFP}) &= \beta_{0,1} + \beta_{1,1}\texttt{year} + \beta_{2,1}\texttt{age}_i + \beta_{3,1}\texttt{gender}_i + (b_{0i,1} + b_{1i,1}\texttt{year}) + \varepsilon_{ij1} \\
y_2 = \log(\texttt{DCP}) &= \beta_{0,2} + \beta_{1,2}\texttt{year} + \beta_{2,2}\texttt{age}_i + \beta_{3,2}\texttt{gender}_i + (b_{0i,2} + b_{1i,2}\texttt{year}) + \varepsilon_{ij2} \\
y_3 = \log(\texttt{L3}) &= \beta_{0,3} + \beta_{1,3}\texttt{year} + \beta_{2,3}\texttt{age}_i + \beta_{3,3}\texttt{gender}_i + (b_{0i,3} + b_{1i,3}\texttt{year}) + \varepsilon_{ij3} \\
b_i &\sim N_6(0, D), \text{ and } \varepsilon_{ijk} \sim N(0, \sigma_k^2) \text{ for } k = 1, 2, 3;
\end{aligned}$$

Event time sub-model for time to tumour recurrence

$$\lambda_i(t) = \lambda_0(t) \exp\left\{\gamma_{v1}\texttt{age}_i + \gamma_{v2}\texttt{gender}_i + W_{2i}(t)\right\}$$

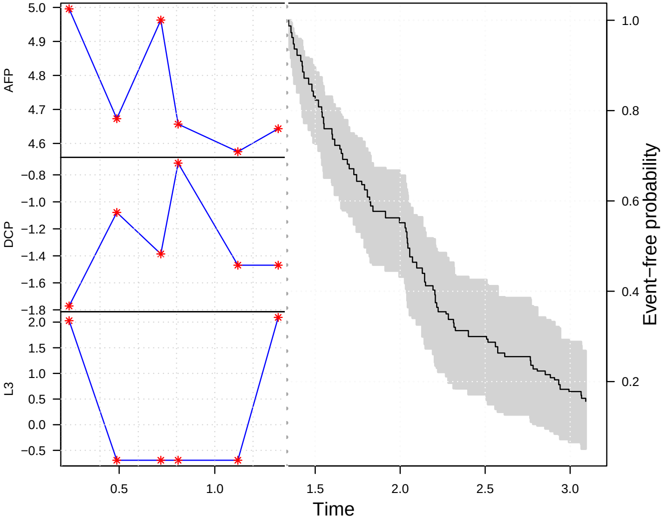
Association structure

$$\begin{aligned}
W_{2i}(t) &= \gamma_{y1} W_{1i}^{(1)}(t) + \gamma_{y2} W_{1i}^{(2)}(t) + \gamma_{y3} W_{1i}^{(3)}(t) \\
&= \gamma_{\texttt{AFP}}(b_{0i,1} + b_{1i,1}\texttt{year}) + \gamma_{\texttt{DCP}}(b_{0i,2} + b_{1i,2}\texttt{year}) + \gamma_{\texttt{L3}}(b_{0i,3} + b_{1i,3}\texttt{year}).
\end{aligned}$$

joineRML::mjoint() code

```
data(HCC)
fit <- mjoint(
formLongFixed = list(
"AFP" = log(AFP) ~ year+age+gender,
"DCP" = log(DCP) ~ year+age+gender,
"L3" = log(L3) ~ year+age+gender),
formLongRandom = list(
"AFP" = ~ year | id,
"DCP" = ~ year | id,
"L3" = ~ year | id),
formSurv = Surv(recurtime, recurstatus) ~ age+gender,
data = HCC,
timeVar = "year",
control = list(tol0 = 0.001, .....))
```

Risk prediction for a new patient, a 65-year-old male



Open challenges and Beyond

Methodology

- Project high-dimensional K biomarkers onto a lower order plane, e.g. variable reduction techniques
- Methods to speed-up estimation
- Alternative association structures
- ...

Application

- Stratify patients based on their risk of recurrence for better targeted therapies
- Better surveillance/personalised follow-up strategies that reduce costs/patient burden
- ...

References

Wei, Greg C. and Tanner, Martin A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* 85(411), pp. 699–704.

Hickey, Graeme L et al. (2018). joineRML: A joint model and software package for time-to-event and multivariate longitudinal outcomes. *BMC Medical Research Methodology* 18(50).

Rizopoulos, Dimitris (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics* 67(3), pp. 819–829.