Fast fitting for joint models of survival and multivariate longitudinal data

James Murray

PhD Viva, 1st May 2024

Supervisor: Dr. Pete Philipson

Outline

A whistle-stop tour...

- Introduction: Joint models, motivation
- "Classic" multivariate joint models
- An approximate EM algorithm
- Flexible joint models
- Justification for approximation used
- Post-hoc analysis; prediction
- Application to PBC
- Discussion; future avenues for research

Background: Joint modelling

Initially arose as a solution to answer analytical challenges in HIV/AIDS research.

Predated by both naïve and "two-stage" methods, both of which don't provide wholly efficient use of available data.

Joint models as they appear in the thesis first given in Wulfsohn & Tsiatis (1997) [1].

Joint models typically consists of (at least) two sub-models linked together by shared random effects.
Models for the longitudinal and survival parts are 'joined' together.

Evolution

Since their first (basic) formal presentation in Wulfsohn & Tsiatis, joint models have been expanded and extended in numerous ways:

- Multivariate case:
  $\hookrightarrow$ Measuring association of ≥ 2 longitudinal responses with survival, accounting for correlations.
- Random effects:
  $\hookrightarrow$ More complex (e.g. splines); more longitudinal sub-models
- Alternative sub-models:
  $\hookrightarrow$ Replacing the Cox PH or the longitudinal model by something more appropriate.

Each serve to increase complexity in the joint model and introduce issues in standard fitting routines; reflected in lack-of software for many cases.

Motivating data

One popular application is to primary biliary cirrhosis (PBC).

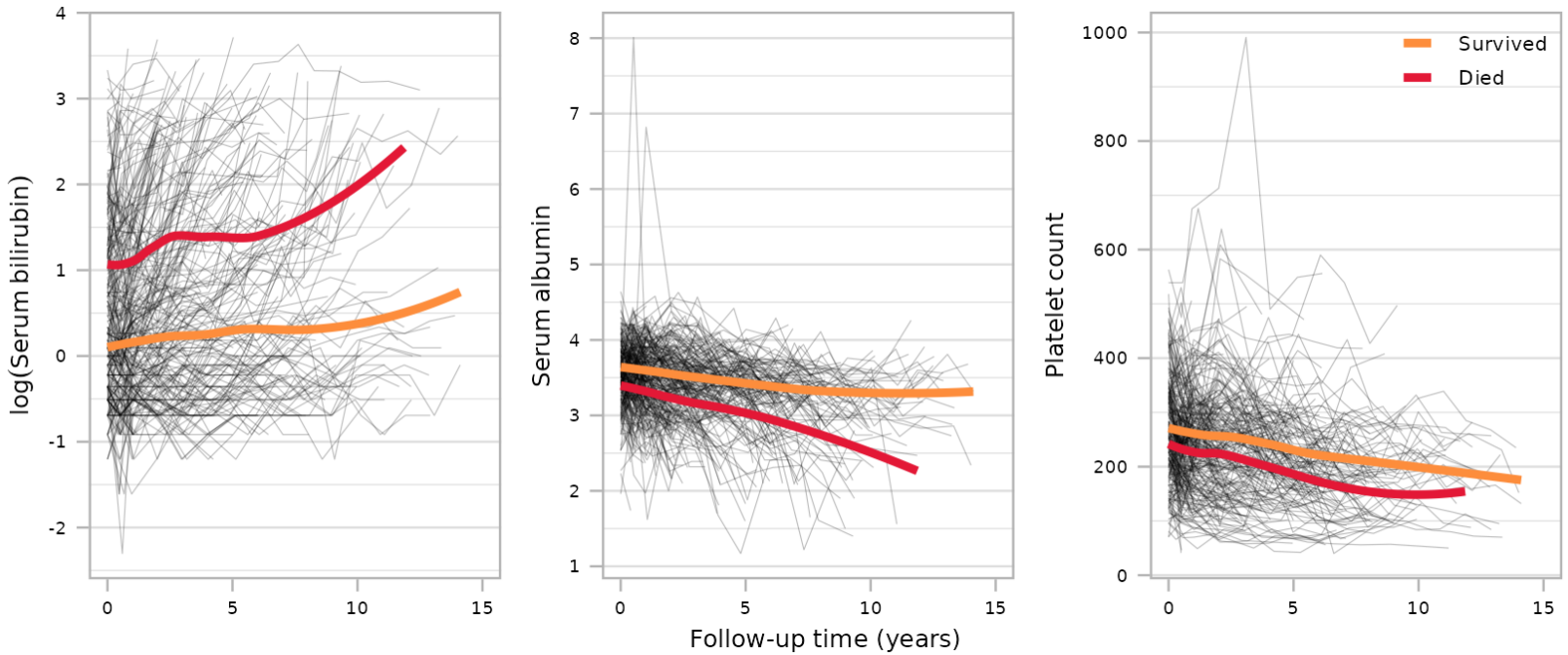


Figure: Patient trajectories for three biomarkers measured during Mayo Clinic trial [2].

Motivating data

One popular application is to primary biliary cirrhosis (PBC).

A joint analysis allows researchers to *simultaneously* answer:

1. How do these biomarkers evolve through time and differ in terms of covariates collected at baseline (e.g. drug allocation, sex, age)?
2. How does the hazard of interest evolve through time an differ in terms of covariates collected at baseline?
3. How is the hazard affected by underlying biomarker values?

Joint modelling framework: Notation

For each $i = 1, \ldots, n$ we observe the $k^{\text{th}}, k = 1, \ldots, K$ longitudinal response $\boldsymbol{Y}_{ik} = (y_{i1k}, \ldots, y_{im_ik})^\top$.

Event time $T_i = \min(T_i^*, C_i)$ and $\Delta_i = 1$ if $T_i^* < C_i$ and $\Delta_i = 0$ otherwise.

Form a joint model by inducing an association between the K longitudinal trajectories and the hazard λ_i:

$$\begin{cases} \boldsymbol{Y}_{ik} & = \mathrm{X}_{ik}(t)\,\boldsymbol{\beta}_k + \mathrm{Z}_{ik}(t)\,\boldsymbol{b}_{ik} + \boldsymbol{\varepsilon}_{ik} \\ & \boldsymbol{b}_{ik} \sim N_{q_k}(0, \mathrm{D}_k), \quad \boldsymbol{\varepsilon}_{ik} \sim N\left(0, \sigma^2_{\varepsilon_k}\right), \quad \boldsymbol{b}_{ik} \perp\!\!\!\perp \boldsymbol{\varepsilon}_{ik} \\ \lambda_i(t) = \lambda_0(t) \exp\left\{ \boldsymbol{S}_i^\top \boldsymbol{\zeta} + \sum_{k=1}^K \gamma_k \boldsymbol{W}_k(t)^\top \boldsymbol{b}_{ik} \right\}. \end{cases}$$

where the association parameter γ_k captures association between $\boldsymbol{b}_{ik}$ on the hazard.

Estimation

Want to estimate the following parameters:

$$\boldsymbol{\Omega} = \left(\text{vech}(\mathrm{D})^\top, \boldsymbol{\beta}^\top, \sigma^2_{\varepsilon_1}, \ldots, \sigma^2_{\varepsilon_K}, \boldsymbol{\gamma}^\top, \boldsymbol{\zeta}^\top\right)^\top$$

which we do my maximising the observed data likelihood via an EM algorithm.

E-step at iteration $(m+1)$:

$$\sum_{i=1}^{n} \mathbb{E}_i \left[\log f\left(\boldsymbol{Y}_i | \boldsymbol{b}_i; \boldsymbol{\Omega}^{(m)}\right) + \log f\left(T_i, \Delta_i | \boldsymbol{b}_i; \boldsymbol{\Omega}^{(m)}\right) + \log f\left(\boldsymbol{b}_i | \boldsymbol{\Omega}^{(m)}\right) \right].$$

calculated against $f\left(\boldsymbol{b}_i | T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega}^{(m)}\right)$

M-step: Formed by maximising *n sets* of conditional expectations.

Estimation

All required expectations necessary in each EM iteration are of form:

$$\mathbb{E}_i[g(\boldsymbol{b}_i)\,|\,T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega}] = \frac{\int_{-\infty}^{\infty} g(\boldsymbol{b}_i)\, f(T_i, \Delta_i|\boldsymbol{b}_i; \boldsymbol{\Omega})\, f(\boldsymbol{b}_i|\boldsymbol{Y}_i; \boldsymbol{\Omega})\, d\boldsymbol{b}_i}{\int_{-\infty}^{\infty} f(T_i, \Delta_i|\boldsymbol{b}_i; \boldsymbol{\Omega})\, f(\boldsymbol{b}_i|\boldsymbol{Y}_i; \boldsymbol{\Omega})\, d\boldsymbol{b}_i}$$

where $f(\boldsymbol{b}_i|\boldsymbol{Y}_i; \boldsymbol{\Omega})$ enjoys tractable form under MVN $\boldsymbol{Y}_i$ [1, 3].

Numerical methods used to evaluate these multi-dimensional integrals.

Main source of computation burden – especially with *more complex* model specifications – potentially precluding uptake.

An approximate EM algorithm

Key issue is that $f(\boldsymbol{b}_i | T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega})$ is potentially high dimensioned.

Bernhardt *et al.* (2015) [4] propose

$$\boldsymbol{b}_i | T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega}^{(m)} \overset{\text{appx.}}{\sim} N\left(\hat{\boldsymbol{b}}_i, \hat{\Sigma}_i\right)$$

Thereby allowing all requisite expectations $\mathbb{E}_i\left[g(\boldsymbol{b}_i) \,|\, T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega}^{(m)}\right]$ to be taken against a univariate normal distribution.

Originally proposed in context of a multivariate joint model with a logistic regression model in place of the Cox PH.

Novel contribution then the extension to more 'traditional' joint models.

An approximate EM algorithm

1. Obtain initial conditions $\boldsymbol{\Omega}^{(0)}$
2. For *any* iteration $(m+1)$:
   - 2.1 Obtain $\hat{\boldsymbol{b}}_i$ and $\hat{\Sigma}_i$ by maximising $\log f\left(\boldsymbol{b}_i, T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega}^{(m)}\right)$ using `optim`
   - 2.2 Use the normal approximation to update the parameter vector $\boldsymbol{\Omega}^{(m)} \rightarrow \boldsymbol{\Omega}^{(m+1)}$.
3. Check for convergence
4. Repeat steps 2. and 3. for at least four iterations, then exit when 3. satisfied.

An approximate EM algorithm: Results

Many simulation studies carried out to ascertain performance.

- Sample size;
- Length of follow-up;
- Number of responses etc.

All studies tabulated and presented graphically.

Good performance across all simulations considered as well as sensitivity analyses.

Compared with `joineRML` [3] where results were very similar.

Non-exponential increase in computation time observed.

Flexible specifications

Gaussian assumption ubiquitous but may not best represent data

$\hookrightarrow$ Important to accommodate (range of) response types

If $\boldsymbol{Y}_i|\boldsymbol{b}_i$ assumed normal, $f(\boldsymbol{b}_i|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega})$ is tractable

Issues arise when this assumption isn't met and the above thereby not tractable.

$\hookrightarrow$ Led to predominantly Bayesian approaches to inference in the literature.

Promising then that the normal approximation eschews consideration of $\boldsymbol{Y}_i|\boldsymbol{b}_i$, instead collapsing down $f(\boldsymbol{b}_i|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega})$ to $N(\hat{\boldsymbol{b}}_i, \hat{\Sigma}_i)$

Flexible specifications

Considered six exponential families for GLMMs:

(Gaussian)	Poisson	Binomial
Negative binomial	Generalised Poisson	Gamma

Estimation of $\boldsymbol{\Omega}$ (potentially with dispersion parameters) via approximate EM algorithm.

Simulations chosen to reflect possible modelling scenarios and investigate performance in 'non-standard' scenarios.

$\hookrightarrow$ Performance good overall, worst performance for binomial.

$\hookrightarrow$ No obvious deterioration in performance in terms of estimation or computation time

Justification for normal approximation

Approximating $\boldsymbol{b}_i | T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega} \sim N\big(\hat{\boldsymbol{b}}_i, \hat{\Sigma}_i\big)$ throughout
$\hookrightarrow$ Good idea to investigate and justify this!

Compared the 'true' posterior for given scenario $f\big(\boldsymbol{b}_i | T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega}^{(\mathrm{TRUE})}\big)$ to $N\big(\hat{\boldsymbol{b}}_i, \hat{\Sigma}_i\big)$ visually and via 'coverage' of the approximation.

Overall the approximation does appear reasonable, but evidence to suggest it's over-confident (i.e. $\hat{\Sigma}_i$ overestimated slightly).

Additionally investigated differences in $N\big(\hat{\boldsymbol{b}}_i, \hat{\Sigma}_i\big)$ when $\hat{\boldsymbol{b}}_i, \hat{\Sigma}_i$ obtained with/out $\{T_i, \Delta_i\}$.

Non-exhaustive, but allow us to get a handle on $N\big(\hat{\boldsymbol{b}}_i, \hat{\Sigma}_i\big)$.

Model diagnostics & prediction

In lieu of a 'joint' residual, consider one for each sub-model.

Hypothesis testing: Wald tests, AIC, BIC.

Dynamic predictions [5], future survival to time u given data up-to t:

- Estimation of $\pi_i(u|t)$ involves $\boldsymbol{b}_i | T_i^* > t, \boldsymbol{Y}_{i1}(t), \ldots, \boldsymbol{Y}_{iK}(t); \hat{\boldsymbol{\Omega}}$. Either by empirical Bayes or Monte Carlo scheme.
- Ascertain predictive performance across 'windows' $w = (T_{\text{start}}, h]$ $\uparrow T_{\text{start}} \implies$ more information available $\implies$ better performance?
- Set out performance measures $\text{AUC}(w)$ and $\widehat{\text{PE}}(w)$
- Comparing nested models
- Correcting for optimism in these estimates.

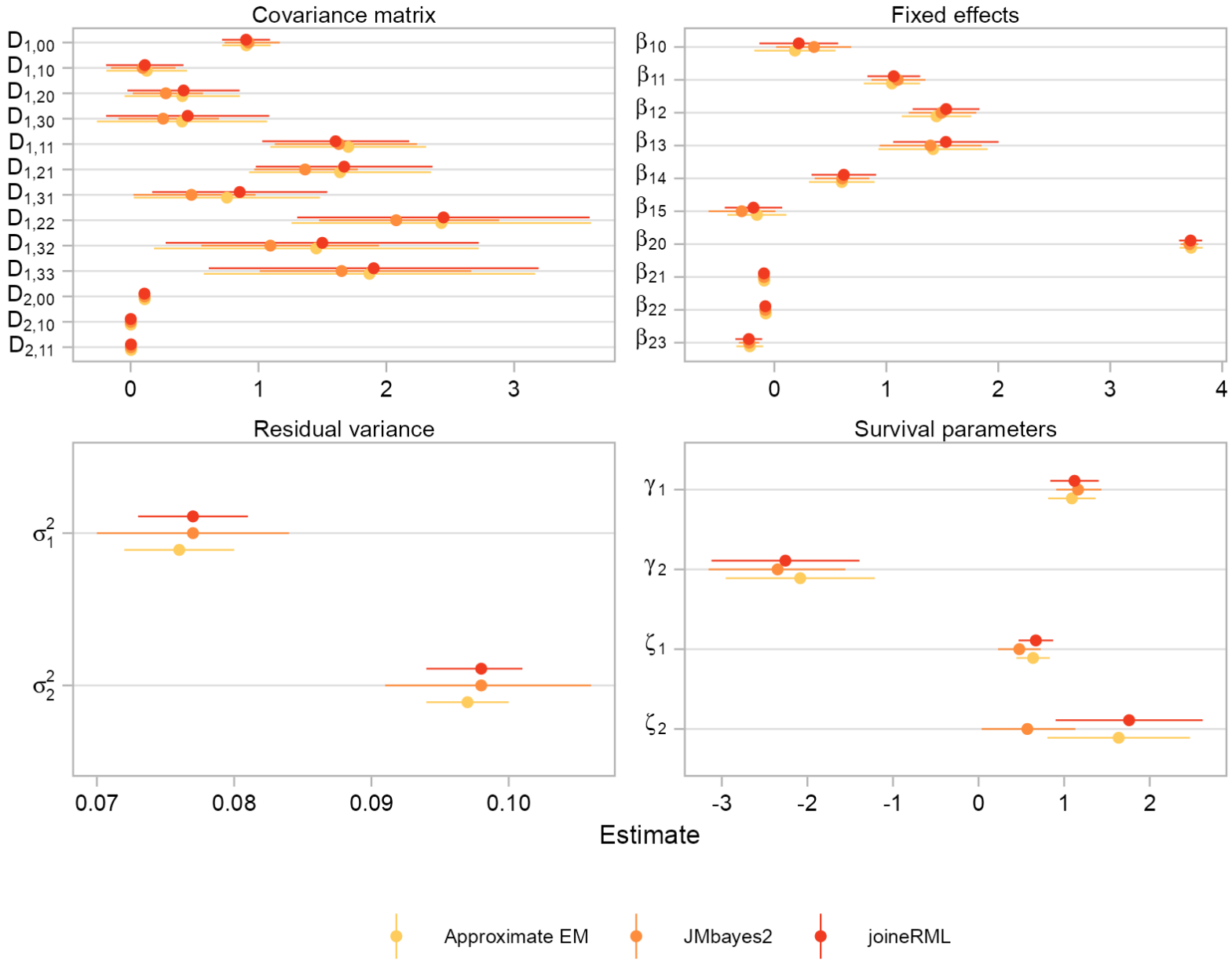
Basically, setting out metrics to use in an application...

Application: Primary biliary cirrhosis

Aim was to present a start-to-finish model building process utilising approximate EM for joint model portion:

- Exploratory analysis and data description:
  Clinical outcome, baseline covariates, candidate longitudinal markers.
- Survival sub-model selection
- Longitudinal sub-model selection for each biomarker:
  Combinations of fixed effects, time specification
- Strength of *univariate* associations
- 'Groups' of markers defining separate multivariate fits
- Trivariate model $\rightarrow$ bivariate model containing serum bilirubin and serum albumin.
- Compared with existing software.

Application: Primary biliary cirrhosis



Conclusions & future work

Vested interest in multivariate joint models:

- Likely better prediction capabilities;
- Use more available information;
- (Series of) univariate fits serves to ignore potential correlations.

However bring with them multidimensional integration, presenting significant computational challenge

Aim was to investigate alternative approaches enabling faster fitting

Repurposed an approximate EM algorithm to lessen computational burden felt in fitting these models

Extended beyond the typically Gaussian paradigm

Conclusions & future work

Avenues for future work

Survival sub-model:

- Competing risks: Patients can experience multiple events
- Accelerated failure time models

Further exponential family members, e.g. :

- Lots of potential count models
- Zero-inflation and zero-truncation

Potential methods for faster computation:

- Linear scan algorithm
- Functional principal components
- Automatic differentiation

Bibliography

1. S. Wulfsohn, M. & A. Tsiatis, A. A Joint Model for Survival and Longitudinal Data Measured with Error. *Biometrics* **53,** 330–339 (1997).
2. Murtaugh, P. A. *et al.* Primary biliary cirrhosis: Prediction of short-term survival based on repeated patient visits. *Hepatology* **20,** 126–134 (1994).
3. G. L. Hickey, P. Philipson, A. Jorgensen & R. Kolamunnage-Dona. joineRML: A joint model and software package for time-to-event and multivariate longitudinal outcomes. *BMC Med. Res. Methodol.* **50** (2018).
4. P. W. Bernhardt, D. Zhang & H. J. Wang. A fast EM Algorithm for Fitting Joint Models of a Binary Response to Multiple Longitudinal Covariates Subject to Detection Limits. *Comput. Stat. Data Anal.* **85,** 37–53 (2015).
5. Rizopoulos, D. Dynamic Predictions and Prospective Accuracy in Joint Models for Longitudinal and Time-to-Event Data. *Biometrics* **67,** 819–829 (2011).