Faster Fitting for Joint Models of Survival and Multivariate Longitudinal Data

James Murray

Thesis submitted for the degree of
Doctor of Philosophy

School of Mathematics, Statistics & Physics
Newcastle University
Newcastle upon Tyne
United Kingdom

May 2024

Abstract

Some twenty-five years after the first introduction of joint modelling, clinical trials across multiple disease areas routinely collect information on many longitudinal biomarkers. This represents opportunities and challenges: Multivariate data likely provides better discrimination capabilities from a prediction standpoint, furthermore disregarding the multivariate nature of the data is tantamount to ignoring potentially informative correlations between these longitudinal trajectories; on the other hand, the multidimensional integrals which arise as part of parameter estimation under traditional approaches present significant computational and statistical difficulties.

We investigate alternative approaches which enable faster fitting of joint models of survival and *multivariate* longitudinal data. An approximate expectation maximisation algorithm relatively dormant in the literature is repurposed to lessen the computational burden felt by traditional joint models, leading to faster fitting. Furthermore, we extend beyond the typically-used restrictive longitudinal specifications in such models in favour of more flexible, potentially complex, specifications.

Extensive simulation studies are carried out, which establish good parameter estimation capabilities of the proposed algorithm under many scenarios. Additionally, a thorough application in the disease area of cirrhosis is carried out, with the algorithm used throughout in building a complex joint model. In both the simulation studies and application, we noted high levels of agreement with established methodology, with the algorithm demonstrating faster computation times.

Acknowledgements

I am enormously grateful to my supervisor, Pete Philipson, for his unwavering support, empathy, and patience over my studies. His guidance helped me develop as a researcher, and I'll miss our weekly meetings filled with debugging, rewriting and having a good natter. Additionally, Pete was immensely supportive through a period of uncertainty in my life, offering great advice and remaining a stalwart source of encouragement, for which I'm especially thankful.
Many thanks to Dr. Joe Matthews and Dr. Lisa McFetridge for examining this thesis.

I have been lucky enough to make many friends within the department, whether this was through sharing an office or eating lunch together, they made the department a great place to work: Ryan Doran, Thomas Flynn, Nicola Hewett, Sam Hartharn-Evans, Nick Keepfer, Kieran Peel and Tom Billam. I'd also like to thank the fantastic computing officer Michael Beaty, who the School is lucky to have, for putting up with many inane queries from me. Outside of the department I'd like to thank my friends Eve and Aaron, Charlotte, as well as Anna, Sophie, India, and Laura. Last, but obviously not least, I'm forever grateful to my Mum and Dad for their unwavering support and encouragement throughout my journey in education, for looking after Beans, and much more.

I would also like to thank the Engineering and Physical Sciences Research Council for funding my PhD studies (grant reference EP/V520184/1).

Declaration

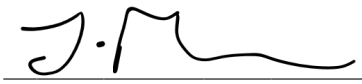
I declare that the work presented in this thesis has been done by myself and has not been submitted for the award of any other academic degree elsewhere. The thesis is supported by background information written by myself, drawing on many existing references in the literature, which have been cited appropriately.

The work presented in Chapter 3 is based on the publication Murray, J., Philipson, P., 2022. *A fast approximate EM algorithm for joint models of survival and multivariate longitudinal data*. Computational Statistics & Data Analysis 170, 107438.

The work presented in Chapter 4, and to a lesser extent Chapter 5, is based on the publication Murray, J., Philipson, P., 2023. *Fast estimation for generalised multivariate joint models using an approximate EM algorithm*. Computational Statistics & Data Analysis 187, 107819.

The work presented in Chapters 1 and 2 borrows small parts from both of the above publications.

Word count for this thesis: 36,838.



James Murray

Contents

Chapter 1

Introduction

1.1 Background to the joint modelling of survival and longitudinal data

In many areas of health research, collection of (potentially many) repeated measurements which are censored by a terminal event are commonplace. Interest then falls on the longitudinal trajectories, risk of event occurrence, and the relationship between the two. The longitudinal process is endogenous and time-dependent if thought to be related to the risk of event occurrence, but simply including the process in de facto modelling approaches is inappropriate (Prentice, 1982; Kalbfleisch and Prentice, 2002; Sweeting and Thompson, 2011).

When the longitudinal outcome and the time-to-event of interest are related, the modelling process must consider dropouts as non-random. Incorporation of the survival information into the longitudinal process has an equivalence to taking into account the effect of an informative missing data process (Sweeting and Thompson, 2011). Conversely, incorporation of the longitudinal information into the survival process improves the fit of regression (i.e. reducing biases arising from separate model fits) and importantly provides inference on the relation between the event time and the longitudinal process (Wulfsohn and Tsiatis, 1997; Sweeting and Thompson, 2011).

Joint modelling arose as a solution to problems arising from HIV/AIDS research in the first instance. Namely, it allowed researchers to *simultaneously* answer three questions of scientific interest:

1. How do the biomarker(s) of interest evolve through time and differ in terms of the other covariates collected at baseline (e.g. drug allocation, sex)?

2. How does the hazard for the event of interest evolve through time and differ in terms of the other covariates collected at baseline?

3. How is this hazard affected by underlying values of the biomarker(s) of interest?

At heart then, a joint model consists of at least two sub-models with some shared random effects that are combined into one larger meta-model by linking these random effects. Commonly adopted sub-models are a linear mixed effects model (LMM, Laird and Ware (1982)), and a Cox proportional hazards (PH) model (Cox, 1972) for the longitudinal and survival components respectively.

Early efforts however were naïve in their implementation, simply including the biomarker as a time-dependent covariate in the PH model. One such example of this approach being Andersen and Gill (1982), producing estimates for association which were underestimated (Prentice, 1982; Sweeting and Thompson, 2011). Additionally, such an approach does not model the longitudinal profile itself, which was one research question we sought to answer.

Next saw so-called 'two-stage' models which provided improvement over the preceding naïve approach. Essentially a 'two-stage' model extracts the random effects from the LMM and models these as time-varying covariates in the PH model. Both Tsiatis et al. (1995) and Gruttola and Tu (1994) used this approach along with an application to aforementioned HIV/AIDS data, albeit with slight reparameterisation of the survival process in the latter example. This modelling approach is relatively easy to implement, and reduced the bias obtained in parameter estimates in comparison with naïve methods (Dafni and Tsiatis, 1998).

Joint models, as they appear in this thesis, then arose in Wulfsohn and Tsiatis (1997). Often seen as the 'first' joint modelling paper, it was in fact predated by efforts conducted under Bayesian hierarchical frameworks, for instance Berzuini and Larizza (1996) and Faucett and Thomas (1996). Given these Bayesian efforts predated standard software such as `WinBUGS` (Spiegelhalter et al., 1999), or because clinicians typically operate under the frequentist paradigm, these are perhaps secondary to the seminal Wulfsohn and Tsiatis (1997) in retrospect.

Wulfsohn and Tsiatis (1997) introduced the joint modelling of longitudinal and time-to-event data using maximum likelihood estimation via the Expectation Maximisation (EM) algorithm (Dempster et al., 1977). Their approach accounted for the informative dropout process and incorporated the observed survival information in modelling the longitudinal process; essentially the available information is most efficiently used (Tsiatis and Davidian, 2004). Justifications for joint models over the naïve and 'two-stage' models abound in literature, two examples showcasing the reduced bias obtained from the joint model are Ibrahim et al. (2010) and Sweeting and Thompson (2011), the latter also highlighting that

joint models perform well despite model misspecification.

1.2 Evolution of joint models

1.2.1 Random effects specification

Wulfsohn and Tsiatis (1997) demonstrated their approach using a simplified model containing only the random effects in both the longitudinal and survival sub-models in an application to the HIV/AIDS data which spurred its initial materialisation. Perhaps acting as something of a 'proof of concept', it neatly set the stage for extensions to the basic model presented therein. Tsiatis et al. (1995), Faucett and Thomas (1996), Bycott and Taylor (1998), Dafni and Tsiatis (1998), and Wulfsohn and Tsiatis (1997) all assumed bivariate Gaussian random effects in the form of a random intercept and slope structure when approaching the joint model. In later literature especially, the random effects have become increasingly complex, with use of e.g. spline terms (Martins, 2022; Rustand et al., 2023). Elsewhere, Henderson et al. (2000) added complexity to the random effects specification by introducing a stationary Gaussian process, similarly carried out in Xu and Zeger (2001) using Bayesian methods. Finally, Wang and Taylor (2001) provide an example of using a non-stationary Gaussian process.

The shared random effects parameterisation was popular earlier in the literature, relating subject-specific characteristics affecting both the longitudinal and survival outcomes, allowing for the estimation of the correlation between them. The interpretation here is that deviations from the estimated population-level longitudinal trajectory drive these associations. The so-called 'current value' parameterisation, however, is a popular alternative. In this approach, the observed value of the longitudinal outcome over time is used as a time-varying covariate in the survival model, linking the hazard to the instantaneous (current) value of the longitudinal measurement. This reflects how the most recent measurement is associated with the hazard. Examples include Chi and Ibrahim (2005), Lin et al. (2002), and Andrinopoulou et al. (2021), with the interested reader referred to Table 1 in Hickey et al. (2016).

1.2.2 Alternative sub-models

Joint models do not solely rely on the previously mentioned linear mixed and Cox proportional hazards models. Two (not necessarily exclusive) ways have emerged in the literature.

The first replaces the PH model with a generalised linear model (GLM). For example,

if occurrence of an event of interest is more important than the *timing* of it, then the PH model is replaced by a binary (i.e. logistic) sub-model. Examples include primary endpoints of successful pregnancy (Horrocks and van Den Heuvel, 2009); survival past a certain period of follow-up (Bernhardt et al., 2015); diagnosis of orthostatic hypertension (Hwang et al., 2011, 2015); and diagnosis of late-life major depressive disorder (Li et al., 2015).

In circumstances where it would be inappropriate to employ a LMM for a longitudinal response, it is instead replaced by a suitable member of the exponential family and modelled by a generalised linear mixed model (GLMM). For instance, if the longitudinal response is scored against a Likert-type scale, (partial) proportional odds models could be used (Li et al., 2010; Alam et al., 2021). A biomarker of interest could simply be the presence/absence of some clinically relevant condition, taking the form of a binary longitudinal response. Examples of this include Choi et al. (2015) who include repeated quality of life measures and Rustand et al. (2023) who monitor presence of malformed blood vessels. Finally, the longitudinal response may be best represented by a count regression model, for example Sunethra and Sooriyarachchi (2018) present a joint model with the number of seizures experienced by epilepsy patients is captured by a Poisson GLMM. Zhu et al. (2018) utilise a zero-inflated Poisson (and generalised Poisson) GLMM in modelling daily cigarette count along with time to study dropout. Both Hickey et al. (2016) and Alsefri et al. (2020) provide good overviews of the usage of GLMM sub-models in joint modelling.

1.2.3 Multivariate joint models

When more than one longitudinal response is believed to be associated with the hazard of the event of interest occurring, harnessing all available information in a single joint model would be advantageous: Simply undertaking several *univariate* joint models is tantamount to ignoring potentially important correlations *between* responses, which could lead to inflated coefficient estimates. Instead, joint models with more than one longitudinal response constructing the longitudinal sub-models (*multivariate* joint models) allow us to simultaneously model *all* information.

In literature, multivariate joint models were originally restricted to methodological developments (Lin et al., 2002): Only recently has software development allowed for routine fitting of joint models with potentially many longitudinal responses. Hickey et al. (2018a), Long and Mills (2018), Andrinopoulou et al. (2020), and Philipson et al. (2020) all employ multivariate joint models (fit under a variety of paradigms), wherein the longitudinal sub-models are exclusively constructed by LMMs. Andrinopoulou and Rizopoulos (2016) utilise a Bayesian approach with shrinkage priors to fit a multivariate joint model, whose longitudinal specification includes GLMMs; Rustand et al. (2023) fit a similarly specified

model using Integrated Nested Laplace Approximations (INLA, Rue et al. (2009)). A recent literature review conducted by Alsefri et al. (2020) found that the vast majority of joint models fit under the Bayesian paradigm are done so by Markov Chain Monte Carlo methods.

Despite opportunities to fit multivariate joint models, a common approach is to use dimension reduction techniques, such as functional principal components regression (Li and Luo, 2017; Li et al., 2021) or partial least squares (Wang et al., 2020). These approaches largely arise where *very many* longitudinal responses exist, such that their implementation is precluded by existing approaches. Indeed, Hickey et al. (2018a) hypothesised that approximate methods may be necessary to tackle issues which arise when rich sub-models are considered.

1.2.4 Software available to fit joint models

Initially, uptake of joint models in e.g. clinical application was slow (Gould et al., 2015). One could argue that this was due to familiarity with long-standing methods (e.g. PH models), despite the proven superiority of joint models, which also shared the same interpretation. Of course, a facet which would prohibit uptake of *any* methodology would be access to (or indeed lack thereof) readily available packages/libraries implemented in popular statistical software.

We briefly note implementation of joint modelling in Statistical software `Stata` through package `stjm` (Crowther et al., 2013), as well as usage of `SAS` to fit joint models through macro `%JM` (Garcia-Hernandez and Rizopoulos, 2018), and continue instead with a focus on `R` (R Core Team, 2020) packages available on the comprehensive R archive network (CRAN, `https://cran.r-project.org/`). Furgal et al. (2019) offer a comprehensive review and comparison via simulation studies of `R` packages `JM` (Rizopoulos, 2010); `joineR` (Philipson et al., 2018) and `JMbayes` (Rizopoulos, 2016), all of which fit *univariate* Gaussian joint models.

Only recently has software become available for the multivariate case. Operating under maximum likelihood, `joineRML` (Hickey et al., 2018a) and under Bayesian methods `JMbayes2` (Rizopoulos et al., 2021), the practitioner is able to readily fit multivariate joint models. Most recently, `INLAjoint` (Rustand et al., 2023) has provided an approximate Bayesian approach to joint modelling via `INLA`.

1.3 Motivating study: Primary biliary cirrhosis

We briefly present publicly available data which which motivates (multivariate) joint modelling. We focus here on data arising from a clinical trial on human subjects. This isn't to say the relationships between longitudinal responses and an event-time of interest are *only* of interest in a clinical setting; they will be of interest in other disciplines too.

Primary biliary cirrhosis (PBC) is a chronic liver disease in which the bile ducts become injured, leading to a build-up of bile in the liver, which can damage it and lead to cirrhosis. If left untreated or otherwise reaches an advanced stage, PBC can lead to severe complications such as liver failure, hypertension and ultimately mortality. The progression of PBC was studied in 312 patients between 1974 and 1984 at the Mayo clinic (Murtaugh et al., 1994). In this study, the patients were randomised and received either placebo or active treatment D-penicillamine. Several biomarkers associated with liver function were repeatedly measured during follow-up, as well as information regarding the time until the first occurrence of either death, receiving a liver transplantation, or the end of study.

The existence of multiple longitudinal biomarkers as well as information regarding a time to event of interest has led to the PBC data becoming a popular example in existing literature (Hickey et al., 2018a; Dai and Pan, 2018; Andrinopoulou and Rizopoulos, 2016; Dil and Karasoy, 2020). The longitudinal trajectories of three biomarkers, (log) serum bilirubin, albumin, and platelet count (i.e. the number of red blood cells) are presented in Figure 1.1 where we distinguish between those who died and those who did not. At a precursory glance, these trajectories could lead one to hypothesise that lower values of albumin and platelet count increase the risk of mortality, with the same true for higher values of serum bilirubin. Here, a joint model would allow us to investigate this relation along with sub-model specific inference as enumerated in Section 1.1

1.4 Thesis outline

With an overview of joint modelling and its development in literature provided in Sections 1.1 and 1.2, we proceed by presenting the structure of the thesis on a chapter-by-chapter basis.

In Chapter 2, we lay the foundations for the thesis by first establishing the notation to be used throughout. We then introduce the multivariate extension of the 'classic' joint models, which are our specific models of interest, elucidating parameters of interest as well as the estimation routine. Couched in a semiparametric maximum likelihood approach we provide an overview of the EM algorithm, aligning our methodology with 'classic' literature. Stemming from this approach, we introduce numerical integration methods.

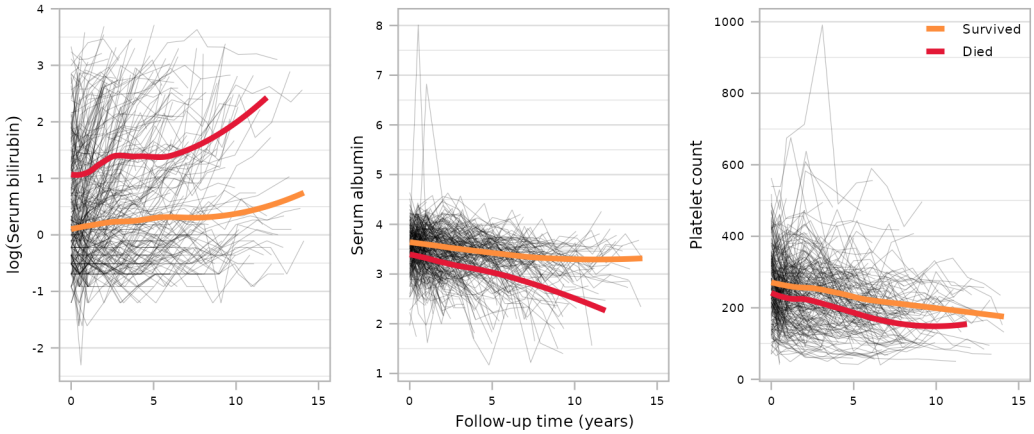


Figure 1.1: Longitudinal trajectories for three chosen biomarkers from PBC data. Grey lines show individual trajectories and overlaid smoothed (LOESS) curves the average trajectories for those who experienced mortality during follow-up and those who did not.

We further outline how one can 'complete' inference and obtain standard errors for a fitted joint model. Finally, we offer a comprehensive overview of simulating data within the joint modeling framework; a crucial technique for the forthcoming chapters.

In Chapter 3, we draw attention to the computational burden which precludes, or makes overtly cumbersome, estimation for multivariate joint models by maximum likelihood. We introduce a normal approximation first proffered by Bernhardt et al. (2015), with details of this approximate EM algorithm provided; the approximation is used exhaustively throughout the thesis. Extensive simulation results are given, which serve to showcase the performance of the algorithm. Additionally, we offer comparison in terms of computation time as well as parameter estimates with existing software which is similar in its estimation approach.

In Chapter 4 we continue utilising the approximate EM algorithm but extend the longitudinal sub-model to the non-Gaussian case; the idea being that many clinical data will not be continuous, or the Gaussian assumption not otherwise appropriate. We consider five different exponential families and provide an overview of the estimation procedure required for these so-called 'generalised' multivariate joint models in a similar spirit to Chapter 3. Once more we provide multitudes of simulation study results to exhibit performance of the approximate EM algorithm.

Owing to its large-scale use in the thesis it is of considerable importance to thoroughly investigate, understand, and justify the normal approximation. In Chapter 5, beginning with its foundations before delineating strategies to justify said approximation, we seek to achieve this; swathes of results are provided along with interpretation.

We then shift focus to avenues for analysis of a fitted joint model. That is, with the

methodology outlined in Chapters 2–4, what can one *do* with the fitted joint model: Does it fit the data well? How can it be used for predictions? Such questions are answered in Chapter 6 where we collate post-hoc analyses, from defining residuals for the two processes, to outlining how one can bridge from a joint model to a predictive one.

Chapter 7 can be viewed as something of an amalgamation of Chapters 2, 3, 4, and 6. We seek to provide an extensive application to the motivating set of clinical data from Section 1.3: The aim of the chapter is to identify, and perform post-hoc analyses on, the 'best-fitting' (multivariate) joint model for the data.

Finally, the thesis is brought to a close by our conclusions along with discussions of possible future research avenues in Chapter 8. The thesis is supplemented by Appendices A–C which house an assortment of extra derivations and results (both tabulated and graphical). Furthermore Appendix D provides a brief overview of the `R` package built alongside the work in Chapter 4, `gmvjoint` (Murray and Philipson, 2023), which is used to obtain *all results* presented in the thesis.

Chapter 2

The Classic Joint Modelling Framework

In this chapter, the multivariate joint modelling framework under which we operate is defined, along with requisite notation. Following this we detail parameter estimation via maximum likelihood and data simulation techniques for the constituent sub-models as well as the joint model itself. We note that we present the '*classic*' joint model in the first instance, named as such since early literature exclusively considered responses which were assumed Gaussian. In Chapter 4 we relax this assumption and introduce a breadth of other response types and in Chapter 6 explore avenues for post-hoc analyses and prediction.

2.1 Joint modelling

2.1.1 Models, likelihood and notation

For each subject $i = 1, \ldots, n$ we observe $\boldsymbol{Y}_i = \left(\boldsymbol{Y}_{i1}{}^\top, \ldots, \boldsymbol{Y}_{iK}{}^\top\right)^\top$ where each $\boldsymbol{Y}_{ik}, k = 1, \ldots, K$ denotes the k^{th} longitudinal response of interest $\boldsymbol{Y}_{ik} = (y_{i1k}, \ldots, y_{im_{ik}k})^\top$. Each of the K responses are measured m_{ik} times, which can differ between subjects and responses. We observe a (possibly right-censored) event time $T_i = \min(T_i^*, C_i)$ where T_i^* denotes the true event time and C_i the independent potential censoring time, subsequently introducing failure indicator Δ_i which is unity if $T_i^* < C_i$ and zero otherwise.

We adopt the following linear mixed effects model for the k^{th} longitudinal response

$$\begin{aligned} \boldsymbol{Y}_{ik} &= \mathrm{X}_{ik}(t)\,\boldsymbol{\beta}_k + \mathrm{Z}_{ik}(t)\,\boldsymbol{b}_{ik} + \boldsymbol{\varepsilon}_{ik} \\ \boldsymbol{b}_{ik} &\sim N_{q_k}(0, \mathrm{D}_k)\,, \qquad \boldsymbol{\varepsilon}_{ik} \sim N\left(0, \sigma^2_{\varepsilon_k}\right), \qquad \boldsymbol{b}_{ik} \perp\!\!\!\perp \boldsymbol{\varepsilon}_{ik}, \end{aligned} \tag{2.1}$$

Here, X_{ik} denotes the (possibly time-dependent) design matrix for the fixed effects of interest to the k^{th} longitudinal response for subject i with corresponding p_k-vector of coefficients $\boldsymbol{\beta}_k$. Likewise, Z_{ik} is a (possibly time-dependent) random effects design matrix and $\boldsymbol{b}_{ik}$ the subject-specific q_k-vector of random effects. These random effects are assumed to follow a zero-mean multivariate normal distribution with covariance matrix D_k. Across the K longitudinal responses, we establish collections of the fixed effects $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_K^\top\right)^\top$ as well as the random effects $\boldsymbol{b}_i = \left(\boldsymbol{b}_{i1}{}^\top, \ldots, \boldsymbol{b}_{iK}{}^\top\right)^\top$ for subject i.

We form the joint model by inducing an association between the K longitudinal trajectories and the hazard λ_i through inclusion of the random effects $\boldsymbol{b}_{ik}$. The sub-model for the event-time process is the usual Cox proportional hazards (PH) model

$$\lambda_i(t) = \lambda_0(t) \exp\left\{ \boldsymbol{S}_i^\top \boldsymbol{\zeta} + \sum_{k=1}^{K} \gamma_k \boldsymbol{W}_k(t)^\top \boldsymbol{b}_{ik} \right\}. \tag{2.2}$$

Here $\boldsymbol{S}_i$ is the vector of baseline covariates of interest to the event-time process and $\boldsymbol{\zeta}$ the corresponding p_s–vector of coefficients. The baseline hazard $\lambda_0(\cdot)$ is treated as a nuisance parameter and left unspecified. Parameter γ_k represents the strength of association between the random effects for the k^{th} longitudinal response in (2.1) and the hazard, with $\boldsymbol{W}_k(t)$ denoting the appropriate vector function of time corresponding to the random effects structure on the k^{th} longitudinal response. This could take the form of an intercept and slope, natural cubic splines and so on. As an example, consider $K = 2$ responses, the first modelled under an intercept and slope specification and the other by random intercept only, the exponent in (2.2) becomes $\exp\left\{ \boldsymbol{S}_i^\top \boldsymbol{\zeta} + \gamma_1 (1, t)^\top (b_{i10} b_{i11}) + \gamma_2 b_{i20} \right\}$. One interprets γ_k as any parameter in a Cox model: Larger values increase the (log) hazard whereas smaller values decrease it. If γ_k is not significantly different from the null, then a separate analysis would suffice.

It is equally – if not more – popular in literature to define the nature of the association in (2.2) by the current value of the k^{th} linear predictor, rather than only its random effects (see e.g. Table 1 in Hickey et al. (2016)), as was discussed in Section 1.2.1. We elect the shared random effects association structure out of preference alone; deviations away from some population mean trajectory being the driving force behind an observed association. More complex association structures could include stationary Gaussian processes (Henderson et al., 2000; Martins, 2022), or current-value-and-slope parameterisations (Rizopoulos and Ghosh, 2011; Rustand et al., 2023).

2.1.2 Likelihood

As we exclusively consider estimation via (semiparametric) maximum likelihood, we first define the log-likelihood of the joint distribution of the set of observed outcomes for subject i, $\{\boldsymbol{Y}_{i1}, \ldots, \boldsymbol{Y}_{iK}, T_i, \Delta_i\}$. In doing so, we assume that each longitudinal process and the survival process are conditional on the random effects $\boldsymbol{b}_i$: The correlation between measurements in the k^{th} longitudinal process as well as the association between this longitudinal process and the event-time process are accounted for by the random effects. We define the joint density as

$$f(T_i, \Delta_i, \boldsymbol{Y}_i | \boldsymbol{b}_i; \boldsymbol{\Omega}) = f(T_i, \Delta_i | \boldsymbol{b}_i; \boldsymbol{\Omega})\, f(\boldsymbol{Y}_i | \boldsymbol{b}_i; \boldsymbol{\Omega}) \tag{2.3}$$

where $f(\boldsymbol{Y}_i | \boldsymbol{b}_i; \boldsymbol{\Omega}) = \prod_{k=1}^{K} f(\boldsymbol{Y}_{ik} | \boldsymbol{b}_{ik}; \boldsymbol{\Omega})$ can be thought of as i.e. the product of the individual probability density functions (pdf) for the K longitudinal processes and $f(T_i, \Delta_i | \boldsymbol{b}_i; \boldsymbol{\Omega})$ denotes the pdf for the event-time process. The vector of parameters is denoted by $\boldsymbol{\Omega}$. We then define the observed data likelihood for subject i as

$$\begin{aligned} f(T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega}) &= \int_{-\infty}^{\infty} f(T_i, \Delta_i, \boldsymbol{Y}_i, \boldsymbol{b}_i; \boldsymbol{\Omega})\, d\boldsymbol{b}_i \\ &= \int_{-\infty}^{\infty} \left[\prod_{k=1}^{K} f(\boldsymbol{Y}_{ik} | \boldsymbol{b}_{ik}; \boldsymbol{\beta}_k, \sigma_k) \right] f(T_i, \Delta_i | \boldsymbol{b}_i; \boldsymbol{\gamma}, \boldsymbol{\zeta})\, f(\boldsymbol{b}_i | \mathrm{D})\, d\boldsymbol{b}_i. \end{aligned} \tag{2.4}$$

where we have integrated out the unobserved random effects. We have additionally introduced $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)^\top$ as the vector of association parameters across the K responses. We now explicitly define the parameter vector $\boldsymbol{\Omega} = \left(\text{vech}(\mathrm{D})^\top, \boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{\zeta}^\top\right)^\top$, where vech$(\cdot)$ denotes the half-vectorisation of its matrix argument, thus returning all unique elements.

Since we seek to maximise the log-likelihood $\ell(\boldsymbol{\Omega}) = \sum_{i=1}^{n} \log f(T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega})$, we present the logarithm of each constituent density in (2.4). The log-likelihood of the random effects density is given by the multivariate normal

$$\log f(\boldsymbol{b}_i | \mathrm{D}) = -\frac{q}{2} \log(2\pi) - \frac{1}{2} \log |\mathrm{D}| - \frac{1}{2} \boldsymbol{b}_i^\top \mathrm{D}^{-1} \boldsymbol{b}_i, \tag{2.5}$$

where q denotes the dimensionality of the random effects covariance matrix D i.e. $q = \sum_{k=1}^{K} q_k$ and $|M|$ denotes the determinant of matrix M. The event-time process has

log-likelihood

$$\log f(T_i, \Delta_i|\boldsymbol{b}_i; \boldsymbol{\gamma}, \boldsymbol{\zeta}) = \Delta_i \log \lambda_0(T_i) + \Delta_i \left[\boldsymbol{S}_i^\top \boldsymbol{\zeta} + \sum_{k=1}^{K} \gamma_k \boldsymbol{W}_k(T_i)^\top \boldsymbol{b}_{ik} \right] - \int_0^{T_i} \lambda_0(u) \exp\left\{ \boldsymbol{S}_i^\top \boldsymbol{\zeta} + \sum_{k=1}^{K} \gamma_k \boldsymbol{W}_k(u)^\top \boldsymbol{b}_{ik} \right\} \mathrm{d}u, \tag{2.6}$$

herein, the vector function of time corresponding to the specification of the k^{th} random effects, $\boldsymbol{W}_k(\cdot)$ introduced in (2.2), is employed with associated failure time T_i, as well as with each survived failure time in the integrand. Notably under an unspecified baseline hazard $\lambda_0(\cdot)$ the quantity in the integrand in (2.6) is only non-zero *at* the observed failure times (Henderson et al., 2000).

Finally, we consider the the log-likelihood of the longitudinal processes. We first exploit the fact that we believe each of the K responses to be (multivariate) normal, conditional on the random effects. For each subject i we construct block-diagonal matrices across the K longitudinal responses for the covariate information, $\mathrm{X}_i = \bigoplus_{k=1}^{K} \mathrm{X}_{ik}$; the random effects structure $\mathrm{Z}_i = \bigoplus_{k=1}^{K} \mathrm{Z}_{ik}$ and error terms $\mathrm{V}_i = \bigoplus_{k=1}^{K} \sigma_{\varepsilon_k}^2 \mathbb{I}_{m_{ik}}$. Here, $\mathbb{I}_x$ denotes the $x \times x$ identity matrix and $\bigoplus$ denotes the direct matrix sum of its arguments. The log-likelihood for the longitudinal responses is then

$$\log f(\boldsymbol{Y}_i|\boldsymbol{b}_i; \boldsymbol{\Omega}) = -\frac{m_i}{2} \log(2\pi) - \frac{1}{2} \log|\mathrm{V}_i| - \frac{1}{2} (\boldsymbol{Y}_i - \boldsymbol{\eta}_i)^\top \mathrm{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\eta}_i), \tag{2.7}$$
$$\boldsymbol{\eta}_i = \mathrm{X}_i \boldsymbol{\beta} - \mathrm{Z}_i \boldsymbol{b}_i,$$

where $m_i = \sum_{k=1}^{K} m_{ik}$.

2.2 Parameter estimation via the EM algorithm

As we discussed in Section 1.1, initial approaches to fitting joint models were in the form of two-stage approaches. Whilst these were simple to implement, they were found in many instances to produce biased parameter estimates (see e.g. Sweeting and Thompson (2011)). We therefore focus on estimation via (semiparametric) maximum likelihood, as was undertaken in seminal underpinning literature (Wulfsohn and Tsiatis, 1997; Henderson et al., 2000). Maximisation of $\ell(\boldsymbol{\Omega}) = \sum_{i=1}^{n} \log f(T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega})$ with respect to $\boldsymbol{\Omega}$ in this frequentist framework is traditionally undertaken by usage of the Expectation Maximisation (EM) algorithm.

2.2.1 A general overview of the EM algorithm

In situations where a procedure for maximum likelihood estimation would be possible if not for the absence of some data, the EM algorithm has arisen as a popular, broadly applicable algorithm that provides an iterative procedure for computing said MLEs. The underlying idea of the EM algorithm is to construct the *complete* data likelihood – which allows for MLEs to be found iteratively – given an *incomplete* data problem (McLachlan and Krishnan, 2008).

Let $f(\boldsymbol{Y} = \boldsymbol{y}; \boldsymbol{\Omega})$ be the probability density function (pdf) of the observed data vector $\boldsymbol{y}$. We henceforth view the observed data $\boldsymbol{y}$ as being incomplete, and a function of the complete data vector $\boldsymbol{x}$; more formally we have mapping from sample spaces $\mathcal{X}$ to $\mathcal{Y}$. We note that the concept of *incomplete* data covers not only the 'conventional' notion of missing data but additionally when some variable(s) would not be observable in practicality.
We introduce next the pdf for the complete data $f_c(\boldsymbol{X} = \boldsymbol{x}; \boldsymbol{\Omega})$, where the subscript c is used to differentiate this from the above pdf for the observed data, and signifies 'complete'. The complete data log-likelihood, if $\boldsymbol{x}$ were observable, is $\ell_c(\boldsymbol{\Omega}) = \log f_c(\boldsymbol{x}; \boldsymbol{\Omega})$. However, instead of observing the complete data $\boldsymbol{x} \in \mathcal{X}$, we observe the incomplete data $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x}) \in \mathcal{Y}$ as detailed previously. It follows that (McLachlan and Krishnan, 2008)

$$f(\boldsymbol{y}; \boldsymbol{\Omega}) = \int_{\mathcal{X}(\boldsymbol{y})} f_c(\boldsymbol{x}; \boldsymbol{\Omega}) d\boldsymbol{x}.$$

The EM algorithm essentially seeks to solve the incomplete data likelihood equation $\partial \ell(\boldsymbol{\Omega}) / \partial \boldsymbol{\Omega} = \boldsymbol{0}$ by instead using the complete data log-likelihood $\ell_c(\boldsymbol{\Omega})$, which in practise is replaced by its conditional expectation on the observed data $\boldsymbol{y}$ at the current estimate of parameter vector $\boldsymbol{\Omega}$, at say iteration m. The EM algorithm consists of two 'steps' per iteration which are eponymous to the approach: The expectation (E-) and maximisation (M-) step. On the $(m+1)$st iteration, the E-step and M-step are defined as:

E-step Calculate

$$Q\Big(\boldsymbol{\Omega}; \boldsymbol{\Omega}^{(m)}\Big) = \mathbb{E}\Big[\ell_c(\boldsymbol{\Omega}) | \boldsymbol{y}; \boldsymbol{\Omega}^{(m)}\Big]. \tag{2.8}$$

M-step Maximise $Q\Big(\boldsymbol{\Omega}; \boldsymbol{\Omega}^{(m)}\Big)$ with respect to $\boldsymbol{\Omega}$, i.e.

$$\boldsymbol{\Omega}^{(m+1)} = \arg\max_{\boldsymbol{\Omega}} Q\Big(\boldsymbol{\Omega}; \boldsymbol{\Omega}^{(m)}\Big). \tag{2.9}$$

To elucidate further; each EM iteration requires conditional expectations of the form $\mathbb{E}[g(\boldsymbol{x}) | \boldsymbol{y}; \boldsymbol{\Omega}]$, where $g(\boldsymbol{x})$ is some function on the complete data required to form the maximum likelihood update in the M-step, which one obtains through e.g. score equations.

2.2.2 The EM algorithm in a joint modelling framework

With the constituent E- and M-steps laid out in (2.8) and (2.9), respectively, we can turn attention toward parameter estimation for the observed data likelihood (2.4).
The set of *observed* data for subject i is constructed by the longitudinal response(s), $\boldsymbol{Y}_i$, the subject's event time T_i and failure indicator Δ_i: $\{\boldsymbol{Y}_i, T_i, \Delta_i\}$. To be explicit, the *complete* data for subject i is $\{\boldsymbol{Y}_i, T_i, \Delta_i, \boldsymbol{b}_i\}$ whereby every element except for the random effects $\boldsymbol{b}_i$ are observed. These random effects are then treated as *missing data* upon implementation of the EM algorithm.

The E-step (2.8) in the context of a joint model at iteration $(m+1)$ is

$$\begin{aligned} Q\Big(\boldsymbol{\Omega}; \boldsymbol{\Omega}^{(m)}\Big) &= \sum_{i=1}^{n} \mathbb{E}_i[\log f(\boldsymbol{Y}_i, T_i, \Delta_i, \boldsymbol{b}_i|\boldsymbol{\Omega})] \\ &= \sum_{i=1}^{n} \int_{-\infty}^{\infty} \{\log f(\boldsymbol{Y}_i, T_i, \Delta_i, \boldsymbol{b}_i|\boldsymbol{\Omega})\}\, f\big(\boldsymbol{b}_i|\boldsymbol{Y}_i, T_i, \Delta_i; \boldsymbol{\Omega}^{(m)}\big) d\boldsymbol{b}_i, \end{aligned}$$

where the integrand provides the contribution to the expected complete data (log) likelihood by subject i. The expectation $\mathbb{E}_i[\cdot]$ is taken with respect to the conditional distribution of the random effects on the observed data at a current set of parameter estimates $f(\boldsymbol{b}_i|\boldsymbol{Y}_i, T_i, \Delta_i; \boldsymbol{\Omega}^{(m)})$. The expected value of the complete data log-likelihood (which forms this E-step) is

$$\begin{aligned} Q\Big(\boldsymbol{\Omega}; \boldsymbol{\Omega}^{(m)}\Big) = \sum_{i=1}^{n} \mathbb{E}_i \Bigg[&\log f\Big(\boldsymbol{Y}_i|\boldsymbol{b}_i; \boldsymbol{\Omega}^{(m)}\Big) + \log f\Big(T_i, \Delta_i|\boldsymbol{b}_i; \boldsymbol{\Omega}^{(m)}\Big) \\ &+ \log f\Big(\boldsymbol{b}_i|\boldsymbol{\Omega}^{(m)}\Big) \Bigg]. \end{aligned} \tag{2.10}$$

The M-step is formed by maximising n *sets* of conditional expectations of the form $\mathbb{E}_i[g(\boldsymbol{b}_i)\,|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega}]$ found in the preceding E-step, where $g(\boldsymbol{b}_i)$ denotes some necessary function of the random effects and the expectation is calculated against the conditional distribution $f\big(\boldsymbol{b}_i|\boldsymbol{Y}_i, T_i, \Delta_i; \boldsymbol{\Omega}^{(m)}\big)$.

The form of the parameter updates in the M-step are widely reported in literature, each illustrating the form required of $g(\boldsymbol{b}_i)$ in the preceding E-step. Here, the updated parameter estimates $\boldsymbol{\Omega}^{(m+1)}$ are given using the 'current' estimates from the previous iteration,

$\boldsymbol{\Omega}^{(m)}$:

$$\begin{aligned}
\boldsymbol{\beta}^{(m+1)} &= \left(\sum_{i=1}^{n} \mathrm{X}_i^\top \mathrm{X}_i\right)^{-1} \left(\sum_{i=1}^{n} \mathrm{X}_i^\top (\boldsymbol{Y}_i - \mathrm{Z}_i \mathbb{E}_i[\boldsymbol{b}_i])\right); \\
\sigma_k^{2(m+1)} &= \frac{1}{\sum_{i=1}^{n} m_{ik}} \sum_{i=1}^{n} \mathbb{E}_i\left[\left(\boldsymbol{Y}_{ik} - \mathrm{X}_{ik}\boldsymbol{\beta}_k{}^{(m)} - \mathrm{Z}_{ik}\boldsymbol{b}_{ik}\right)^\top \left(\boldsymbol{Y}_{ik} - \mathrm{X}_{ik}\boldsymbol{\beta}_k{}^{(m)} - \mathrm{Z}_{ik}\boldsymbol{b}_{ik}\right)\right]; \\
\mathrm{D}^{(m+1)} &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_i\left[\boldsymbol{b}_i \boldsymbol{b}_i^\top\right]; \\
\lambda_0^{(m+1)}(t) &= \frac{\sum_{i=1}^{n} \Delta_i \mathrm{I}(T_i = t)}{\sum_{i=1}^{n} \mathbb{E}_i\left[\exp\left\{\boldsymbol{S}_i^\top \boldsymbol{\zeta}^{(m)} + \sum_{k=1}^{K} \gamma_k^{(m)} \boldsymbol{W}_k(t)^\top \boldsymbol{b}_{ik}\right\}\right] \mathrm{I}(T_i > t)};
\end{aligned} \tag{2.11}$$

where $I(\cdot)$ is the indicator function. The update for the parameters associated with the survival (log) density (2.6), $\boldsymbol{\gamma}$ and $\boldsymbol{\zeta}$, do not exist in closed form (owing to their housing in an exponent), so are maximised iteratively by a one-step Newton Raphson algorithm. We let $\boldsymbol{\Phi} = \left(\boldsymbol{\gamma}^\top, \boldsymbol{\zeta}^\top\right)^\top$ denote the survival parameters, with update from iteration (m) to $(m+1)$ given by

$$\boldsymbol{\Phi}^{(m+1)} = \boldsymbol{\Phi}^{(m)} - \left[\sum_{i=1}^{n} \mathrm{H}_i\left(\boldsymbol{\Phi}^{(m)}\right)\right]^{-1} \left[\sum_{i=1}^{n} s_i\left(\boldsymbol{\Phi}^{(m)}\right)\right], \tag{2.12}$$

where $s_i(\boldsymbol{\Phi})$ is the gradient vector of the conditional expectation of the requisite complete data (profile) log-likelihood with respect to $\boldsymbol{\Phi}$, and $\mathrm{H}_i(\boldsymbol{\Phi})$ the matrix of second derivatives for subject i. We undertake a more in-depth look at the E-step for these joint models in the next section.

2.3 The E-step for joint models

We require expectations of the form

$$\mathbb{E}_i[g(\boldsymbol{b}_i)\,|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega}],$$

which are taken with respect to the density of the random effects conditioned on the observed data at the current set of parameter estimates, with this 'current set' notation temporarily dropped,

$$\mathbb{E}_i[g(\boldsymbol{b}_i)\,|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega}] = \int_{-\infty}^{\infty} g(\boldsymbol{b}_i)\, f(\boldsymbol{b}_i|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega})\, d\boldsymbol{b}_i. \tag{2.13}$$

In the presence of the random effects $\boldsymbol{b}_i$ the longitudinal process(es) $\boldsymbol{Y}_i$ *do not* inform the survival process $\mathcal{S}_i = \{T_i, \Delta_i\}$; these are conditionally independent given $\boldsymbol{b}_i$. With this in mind we then slightly rewrite our conditional density $f(\boldsymbol{b}_i|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega})$ to be

$$f(\boldsymbol{b}_i|\mathcal{S}_i; \boldsymbol{Y}_i, \boldsymbol{\Omega})$$

where the set of survival information $\mathcal{S}_i$ is momentarily used for ease of presentation, and $\boldsymbol{Y}_i$ is now treated as an extraneous element.

We then proceed using Bayes' theorem to rewrite this conditional density

$$f(\boldsymbol{b}_i|\mathcal{S}_i; \boldsymbol{Y}_i, \boldsymbol{\Omega}) = \frac{f(\mathcal{S}_i|\boldsymbol{b}_i; \boldsymbol{Y}_i, \boldsymbol{\Omega})\, f(\boldsymbol{b}_i|\boldsymbol{Y}_i, \boldsymbol{\Omega})}{f(\mathcal{S}_i|\boldsymbol{Y}_i, \boldsymbol{\Omega})}, \tag{2.14}$$

where the quantity $f(\boldsymbol{b}_i|\boldsymbol{Y}_i, \boldsymbol{\Omega})$ is given in terms of the conditional and marginal probabilities

$$f(\boldsymbol{b}_i|\boldsymbol{Y}_i, \boldsymbol{\Omega}) = \frac{f(\boldsymbol{Y}_i, \boldsymbol{b}_i|\boldsymbol{\Omega})}{f(\boldsymbol{Y}_i|\boldsymbol{\Omega})} = \frac{f(\boldsymbol{Y}_i, \boldsymbol{b}_i|\boldsymbol{\Omega})}{\int_{-\infty}^{\infty} f(\boldsymbol{Y}_i, \boldsymbol{b}_i|\boldsymbol{\Omega})\, d\boldsymbol{b}_i} \equiv \frac{f(\boldsymbol{Y}_i|\boldsymbol{b}_i; \boldsymbol{\Omega})\, f(\boldsymbol{b}_i|\boldsymbol{\Omega})}{\int_{-\infty}^{\infty} f(\boldsymbol{Y}_i|\boldsymbol{b}_i; \boldsymbol{\Omega})\, f(\boldsymbol{b}_i|\boldsymbol{\Omega})\, d\boldsymbol{b}_i}, \tag{2.15}$$

Rewriting the denominator in (2.14) with effort taken to remove the conditioning of $\mathcal{S}_i$ on $\boldsymbol{Y}_i$ we obtain

$$f(\mathcal{S}_i|\boldsymbol{Y}_i, \boldsymbol{\Omega}) = \int_{-\infty}^{\infty} f(\mathcal{S}_i|\boldsymbol{b}_i; \boldsymbol{\Omega})\, f(\boldsymbol{b}_i|\boldsymbol{Y}_i, \boldsymbol{\Omega})\, d\boldsymbol{b}_i.$$

Thus, the density against which expectations are evaluated is (returning to our previously-used notation)

$$f(\boldsymbol{b}_i|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega}) = \frac{f(T_i, \Delta_i|\boldsymbol{b}_i; \boldsymbol{\Omega})\, f(\boldsymbol{b}_i|\boldsymbol{Y}_i; \boldsymbol{\Omega})}{\int_{-\infty}^{\infty} f(T_i, \Delta_i|\boldsymbol{b}_i; \boldsymbol{\Omega})\, f(\boldsymbol{b}_i|\boldsymbol{Y}_i; \boldsymbol{\Omega})\, d\boldsymbol{b}_i}, \tag{2.16}$$

with all required expectations then taking the form

$$\begin{aligned}\mathbb{E}_i[g(\boldsymbol{b}_i)\,|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega}] = {} & \frac{1}{\int_{-\infty}^{\infty} f(T_i, \Delta_i|\boldsymbol{b}_i; \boldsymbol{\Omega})\, f(\boldsymbol{b}_i|\boldsymbol{Y}_i; \boldsymbol{\Omega})\, d\boldsymbol{b}_i} \\ & \times \int_{-\infty}^{\infty} g(\boldsymbol{b}_i)\, f(T_i, \Delta_i|\boldsymbol{b}_i; \boldsymbol{\Omega})\, f(\boldsymbol{b}_i|\boldsymbol{Y}_i; \boldsymbol{\Omega})\, d\boldsymbol{b}_i.\end{aligned} \tag{2.17}$$

For the case when the conditional distribution of $\boldsymbol{Y}_i|\boldsymbol{b}_i$ is assumed to be multivariate normal i.e. $\boldsymbol{Y}_i|\boldsymbol{b}_i \sim N(\mathrm{X}_i\boldsymbol{\beta} + \mathrm{Z}_i\boldsymbol{b}_i, \mathrm{V}_i)$, the density $f(\boldsymbol{b}_i|\boldsymbol{Y}_i; \boldsymbol{\Omega})$ present in (2.14) and, by extension, (2.17) is given by multivariate normal theory (Wulfsohn and Tsiatis, 1997; Lin

et al., 2002; Hickey et al., 2018a)

$$\boldsymbol{b}_i|\boldsymbol{Y}_i;\boldsymbol{\Omega} \sim N\Big(\mathrm{A}_i\Big\{\mathrm{Z}_i^\top \mathrm{V}_i^{-1}(\boldsymbol{Y}_i - \mathrm{X}_i\boldsymbol{\beta})\Big\}, \mathrm{A}_i\Big), \tag{2.18}$$

where $\mathrm{A}_i = \big(\mathrm{Z}_i^\top \mathrm{V}_i^{-1}\mathrm{Z}_i + \mathrm{D}^{-1}\big)^{-1}$. This tractable expression for $f(\boldsymbol{b}_i|\boldsymbol{Y}_i;\boldsymbol{\Omega})$ allows us to avoid the computational burden of evaluating (2.15).

2.3.1 Numerical integration

Fitting joint models can be computationally burdensome, owing in large part to the integral taken over the random effects in calculation of the necessary conditional expectations (2.13), which are necessary in obtaining the parameter estimates at each subsequent M-step (2.11).

This integral does not have an analytic solution; a numerical approach is typically employed to approximate these integrals for each subject. Such evaluation over the random effects serves as the main source of computational woe, and is the main bottleneck in the majority of cases. The random effects are likely not one-dimensional, and computational demand is commensurate with their dimensionality (Philipson et al., 2020).

Several methods for approximation of the conditional expectation (2.13) are found in the literature. Early methods included low-dimensional Gauss-Hermite quadrature or Monte Carlo methods (Wulfsohn and Tsiatis, 1997; Henderson et al., 2000); both since becoming routinely used in available software. More recently, adaptive quadrature methods have been employed, along with Laplace approximations. In the following sections, we seek to outline how these methods are used in the E-step of a joint model to evaluate $\mathbb{E}_i[g(\boldsymbol{b}_i)\,|T_i, \Delta_i, \boldsymbol{Y}_i;\boldsymbol{\Omega}]$.

2.3.2 Gauss-Hermite quadrature

Gauss-Hermite quadrature is a method for approximating definite integrals of the form

$$\int_{-\infty}^{\infty} f(x)\exp\{-x^2\}\,dx.$$

Since the integrand above resembles the normal distribution, Gauss-Hermite quadrature is particularly useful for approximating integrals resembling the normal kernel. Considering our expectation of interest, which we restate from (2.13)

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

under 'standard' Gauss-Hermite quadrature rules we evaluate the above by the sum of ϱ weighted evaluations of the integrand at pre-specified abscissae. The weights $\boldsymbol{w}$ and abcissae $\boldsymbol{v}$ are available in numerous locations, including the `R` package `statmod` (Smyth, 2005). The integral is approximated then by

$$\int_{-\infty}^{\infty} g(\boldsymbol{b}_i) f(\boldsymbol{b}_i|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega})\, d\boldsymbol{b}_i \approx$$
$$2^{q/2} \sum_{l_1=1}^{\varrho} \cdots \sum_{l_q=1}^{\varrho} w_l g\Big(\boldsymbol{v}_l\sqrt{2}\Big)\, f\Big(\boldsymbol{v}_l\sqrt{2}|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega}\Big) \exp\big\{||\boldsymbol{v}_l||^2\big\}, \tag{2.19}$$

where $||x||$ denotes the euclidean norm of its argument. In (2.19) above, the abscissae $\boldsymbol{v}_l = \big(v_{l_1}, \ldots, v_{l_q}\big)^\top$ each have corresponding weights $\boldsymbol{w} = (w_1, \ldots, w_\varrho)^\top$. That is, we extend a univariate Gaussian quadrature rule to the q-dimensioned case (Rizopoulos, 2012b).

Owing to how the specific values of the nodes and weights are assigned to best interpolate a polynomial of degree $2\varrho - 1$, the closeness of the approximation (2.19) to the exact value improves as ϱ increases. Typically, this 'standard' Gauss-Hermite method assumes zero-centering with the shape of a normal distribution. Therefore, another important factor that affects the quality of the approximation is the location of the quadrature points with respect to the modal mass of the integrand. That is, the approximation (2.19) will be poor if most of the mass of $g(\boldsymbol{b}_i) f(\boldsymbol{b}_i|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega})$ lies away from zero, and/or the spread of this integrand is different from the weight function. In the context of (generalised) linear mixed models, Stringer and Bilodeau (2022) argue that this 'shifting' of mass occurs with some regularity.

This 'standard' Gauss-Hermite quadrature provides then a method to (essentially, exactly with large enough ϱ) appraise integrals whose analytic solution is either non-existent, or irritating, to calculate. We draw attention to two main drawbacks to the approach. First, the computational burden of the dimension of the random effects coupled with requirement of a potentially high number of quadrature points being necessary to achieve an approximation of sufficient accuracy. Secondly, one will obtain a poor approximation if the mass of the objective function is located away from zero, and so a more reliable method may be desired.

Indeed, in much of the early literature surrounding joint models, strides were taken away from this numerical integration method in favour of Monte Carlo methods (which we go on to discuss in Section 2.3.5) perhaps because of these issues. In the next section, we examine a method which attempts to circumvent possible malposition of the quadrature routine.

2.3.3 Adaptive Gauss-Hermite quadrature

With consternation surrounding the position of quadrature nodes we outlined in the previous section in mind we explore now an 'adaptive' quadrature rule (Pinheiro and Bates, 1995). This aims to appropriately center and scale abscissae used in evaluation of the integrand (2.13). The integral is approximated by

$$\int_{-\infty}^{\infty} g(\boldsymbol{b}_i) f(\boldsymbol{b}_i|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega})\, d\boldsymbol{b}_i \approx$$
$$2^{q/2}|\hat{\mathrm{C}}_i| \sum_{l_1=1}^{\varrho} \cdots \sum_{l_q=1}^{\varrho} w_l g(\hat{\boldsymbol{v}}_l)\, f(\hat{\boldsymbol{v}}_l|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega}) \exp\{||\boldsymbol{v}_l||^2\}, \tag{2.20}$$

where the aforementioned shift in abcissae is performed by introduction of the term $\hat{\boldsymbol{v}}_l = \hat{\boldsymbol{b}}_i + \sqrt{2}\hat{\mathrm{C}}_i \boldsymbol{v}_l$, with $\boldsymbol{v}_l$ previously defined in addition to

$$\hat{\boldsymbol{b}}_i = \mathrm{argmax}_{\boldsymbol{b}_i}\{\log f(\boldsymbol{Y}_i, T_i, \Delta_i, \boldsymbol{b}_i; \boldsymbol{\Omega})\}, \tag{2.21}$$

$$\hat{\mathrm{H}}_i = -\frac{\partial^2 \log f(\boldsymbol{Y}_i, T_i, \Delta_i, \boldsymbol{b}_i; \boldsymbol{\Omega})}{\partial \boldsymbol{b}_i \partial \boldsymbol{b}_i^\top}\bigg|_{\boldsymbol{b}_i = \hat{\boldsymbol{b}}_i}. \tag{2.22}$$

Finally, the term $\hat{\mathrm{C}}_i$ present in (2.20) is the lower triangular Choleski factorisation which satisfies $\hat{\mathrm{C}}_i \hat{\mathrm{C}}_i^\top = \hat{\mathrm{H}}_i^{-1}$. Rizopoulos (2012b) draw attention to the transformed integrand resembling a $N\big(\mathbf{0}, \frac{1}{2}\mathbb{I}_q\big)$ distribution; proportional to the weight function of Gauss-Hermite quadrature.

Since the evaluation (2.20) centers the abscissae $\boldsymbol{v}$ at the mode $\hat{\boldsymbol{b}}_i$ and scales by the curvature $\hat{\mathrm{H}}_i$ of the complete data (log) likelihood for each of the n subjects, it generally provides a very good approximation of (2.13) (Stringer and Bilodeau, 2022). Typically, this adaptive quadrature rule requires fewer abscissae than its 'standard' counterpart explored in Section 2.3.2 (Rizopoulos, 2012b), owing to its eponymous 'adaptive' quality.

Briefly, on the topic of the number of abscissae, we note little guidance within joint modelling literature for a proposed value for ϱ; simply that increasing it will likely stabilise parameter estimates. Couched in a GLMM setting, Stringer and Bilodeau (2022) recommend setting the number of quadrature points based on two characteristics of the data: The number of 'groups' in the data (e.g. the number of subjects n) and the smallest number of observations across all n groups $M = \min(m_i)$

$$\varrho_{\text{Stringer}} = \left\lceil \frac{3}{2} \log_M n - 2 \right\rceil. \tag{2.23}$$

2.3.4 Pseudo-adaptive Gauss-Hermite quadrature

Belying the promise of greater accuracy and fewer requisite abscissae in the routine outlined in (2.20), large computational cost can manifest due to necessary *repeated* calculation of $\hat{\boldsymbol{b}}_i$ and subsequently $\hat{\mathrm{H}}_i$, given in (2.21) and (2.22), respectively. Explicitly, since (2.20) is housed within the EM algorithm, both $\hat{\boldsymbol{b}}_i$ and $\hat{\mathrm{H}}_i, i = 1, \ldots, n$ need to be obtained at *each* iteration, prior to whatever (potentially expensive) quadrature routine is undertaken.

To circumvent the repeated obtention of n sets of this vector mode and its Hessian matrix, Rizopoulos (2012a) highlight that the (log) density $\log f(\boldsymbol{Y}_i, T_i, \Delta_i; \boldsymbol{\Omega})$ is dominated by the (log) density associated with the linear mixed effects model $\log f(\boldsymbol{Y}_i|\boldsymbol{b}_i; \boldsymbol{\Omega})$, which resembles the shape of some multivariate normal distribution. They argue then that the centering and scaling of abscissae for each individual need only be carried out once, prior to commencement of any EM algorithm. In a (very) similar fashion to (2.20), we approximate

$$\int_{-\infty}^{\infty} g(\boldsymbol{b}_i) f(\boldsymbol{b}_i|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega})\, d\boldsymbol{b}_i \approx$$
$$2^{q/2}|\tilde{\mathrm{C}}_i| \sum_{l_1=1}^{\varrho} \cdots \sum_{l_q=1}^{\varrho} w_l g(\tilde{\boldsymbol{v}}_l)\, f(\tilde{\boldsymbol{v}}_l|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega}) \exp\{||\boldsymbol{v}_l||^2\}\,. \tag{2.24}$$

Here, $\tilde{\boldsymbol{v}}_l = \tilde{\boldsymbol{b}}_i + \sqrt{2}\tilde{\mathrm{C}}_i \boldsymbol{v}_l$ and $\tilde{\mathrm{C}}_i$ is the lower Choleski of $\tilde{\mathrm{H}}_i^{-1}$, where

$$\tilde{\boldsymbol{b}}_i = \mathrm{argmax}_{\boldsymbol{b}_i}\{\log f(\boldsymbol{Y}_i, \boldsymbol{b}_i; \boldsymbol{\Omega})\}\,, \tag{2.25}$$

and $\tilde{\mathrm{H}}_i$ subsequently defined in a corresponding way to (2.22), with the log density $\log f(\boldsymbol{Y}_i, \boldsymbol{b}_i; \boldsymbol{\Omega})$ at mode $\tilde{\boldsymbol{b}}_i$. Given the aforementioned domination of the complete data log-density by the log-density contributed by the linear mixed model, Rizopoulos (2012a) propose, under general regularity conditions,

$$\log f(\boldsymbol{b}_i|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega}) \xrightarrow{\mathrm{P}} N\left(\tilde{\boldsymbol{b}}_i, \tilde{\mathrm{H}}_i^{-1}\right). \tag{2.26}$$

With this 'pseudo-adaptive' method laid out, an immediately obvious advantage is the elusion of repeated (re-)calculation of abscissae for the sample at each EM iteration; compounded by the ease at which *both* $\tilde{\boldsymbol{b}}_i$ *and* $\tilde{\mathrm{H}}_i$, $\forall\ i = 1, \ldots, n$ can be obtained by readily-available software (e.g. `nlme` (Pinheiro et al., 2021), `lme4` (Bates et al., 2015)), with such routines very often used in obtaining initial conditions for $\boldsymbol{\Omega}$. This, along with the 'adaptive' routine outlined in Section 2.3.3, which acts as the foundation for *this* routine, necessitating fewer quadrature points than the 'standard' (Section 2.3.2), results in this 'pseudo-adaptive' quadrature method an attractive avenue.

We note that although (2.20) and (2.24) reduce the number of abscissae required to reli-

ably approximate the integral, the number of evaluations they necessitate still increases exponentially with the dimension of the integrand determined by the number of random effects q.

2.3.5 Monte Carlo methods

Monte Carlo integration is perhaps an obvious alternative to these quadrature methods, as it provides a probabilistic representation of the integral (2.13) (Lemieux, 2009). In Section 2.3, we showed that the conditional expectation $\mathbb{E}_i[g(\boldsymbol{b}_i)\,|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega}]$ can be written as

$$\mathbb{E}_i[g(\boldsymbol{b}_i)\,|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega}] = \frac{\int_{-\infty}^{\infty} g(\boldsymbol{b}_i)\, f(T_i, \Delta_i|\boldsymbol{b}_i; \boldsymbol{\Omega})\, f(\boldsymbol{b}_i|\boldsymbol{Y}_i; \boldsymbol{\Omega})\, d\boldsymbol{b}_i}{\int_{-\infty}^{\infty} f(T_i, \Delta_i|\boldsymbol{b}_i; \boldsymbol{\Omega})\, f(\boldsymbol{b}_i|\boldsymbol{Y}_i; \boldsymbol{\Omega})\, d\boldsymbol{b}_i}. \tag{2.27}$$

where the density $f(T_i, \Delta_i|\boldsymbol{b}_i; \boldsymbol{\Omega})$ is known from (2.6) and, operating under an assumption of multivariate normality in $\boldsymbol{Y}_i|\boldsymbol{b}_i$, the density $f(\boldsymbol{b}_i|\boldsymbol{Y}_i; \boldsymbol{\Omega})$ can be calculated from multivariate normal theory, as given in (2.18).

Monte Carlo integration can therefore be used to approximate (2.27) by first sampling N realizations from (2.18) for each subject and then calculating the ratio of sample mean values for the numerator and denominator integrands present in (2.27). That is,

$$\mathbb{E}_i[g(\boldsymbol{b}_i)\,|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega}] \approx \frac{\frac{1}{N}\sum_{l=1}^{N} g\left(\boldsymbol{b}_i^{(l)}\right) f\left(T_i, \Delta_i|\boldsymbol{b}_i^{(l)}; \boldsymbol{\Omega}\right)}{\frac{1}{N}\sum_{l=1}^{N} f\left(T_i, \Delta_i|\boldsymbol{b}_i^{(l)}; \boldsymbol{\Omega}\right)}. \tag{2.28}$$

To be explicit, the draws $\boldsymbol{b}_i^{(l)},\ l = 1, \ldots, N$ are sampled from the multivariate normal distribution given in (2.18) using, for instance, `rmvnorm` available in package `mvtnorm` (Genz et al., 2021) or `mvrnorm` from package `MASS` (Venables and Ripley, 2002).

Henderson et al. (2000) utilised antithetic Monte Carlo sampling to halve the number of samples required, producing N negatively correlated sample pairs; the negative correlation between these pairs leading to smaller variance in the sample means in (2.28). It is worth mentioning that there is little guidance available with respect to a tactful choice of N. Philipson et al. (2020) outline in greater detail these properties, and additionally draw attention to quasi-Monte Carlo methods for sampling of $\boldsymbol{b}_i|\boldsymbol{Y}_i; \boldsymbol{\Omega}$ couched within the E-step of a joint model.

Notably, Monte Carlo methods do not suffer the same curse of dimensionality as quadrature: The integral approximation given in (2.28) does *not* directly depend on q. However, the random samples from the distribution of $\boldsymbol{b}_i$ may themselves become computationally prohibitive at larger q.

2.4 Calculation of standard errors

With our maximum likelihood estimates $\hat{\boldsymbol{\Omega}}$ obtained from an EM algorithm as delineated in Sections 2.2 and 2.3, we now turn attention to obtaining standard errors (SEs) for the MLEs in order to 'complete' inference for the fitted joint model. Doing so allows us the form the usual Wald confidence intervals for the parameter estimates, obtain p-values, and so on. Generally, standard errors are found by inversion of the information matrix, $\mathcal{I}\big(\hat{\boldsymbol{\Omega}}\big)$.

The lack of *automatic* provision of $\mathcal{I}$ by the EM algorithm lead to several proffered methods to calculate, or approximate, the *observed* information matrix, $\mathcal{I}_o\big(\hat{\boldsymbol{\Omega}}\big)$. Instead of surveying the general-case methods laid out in McLachlan and Krishnan (2008), we focus here on outlining methods for estimating standard errors using the EM algorithm for the joint model specifically.

Hsieh et al. (2006) highlight the major challenge with standard error estimation for joint models is presence of the unspecified (i.e. non-parametric) baseline hazard $\lambda_0(\cdot)$, so the standard asymptotic properties for the MLE $\hat{\boldsymbol{\Omega}}$ afforded by the EM algorithm do not apply. In Section 2.2.2 we therefore essentially considered the *profile* likelihood in parameter updates. This profile likelihood with nonparametric MLE $\hat{\lambda}_0$ shouldn't be dependent upon λ_0. However, Hsieh et al. (2006) admonish that this is *not* the case, since the estimate for $\lambda_0(\cdot)$ appearing in (2.11) holds implicit membership in $\boldsymbol{\Omega}$ upon which the distribution of the random effects is conditionally dependent in (2.17) i.e. itself involves the density $f(T_i, \Delta_i|\boldsymbol{b}_i; \boldsymbol{\gamma}, \boldsymbol{\zeta}, \lambda_0)$.

In light of the challenges stemming from the unspecified λ_0, we present three methods for calculation of SEs. We explore the bootstrap, and two methods which attempt to approximate $\mathcal{I}_o\big(\hat{\boldsymbol{\Omega}}\big)$ in the next sections. We do not compute the standard errors for the baseline hazard $\hat{\lambda}_0$ in all simulations and applications in proceeding chapters: Its high dimensionality, determined by the number of unique event times, could lead to instability in any obtained information matrix, and, moreover, they are likely not of interest considering λ_0 is set as a nuisance parameter at the outset.

2.4.1 The bootstrap

In a scenario where establishing a handle on uncertainty in parameter estimates is supposedly not clear or straight-forward, bootstrapping is perhaps an obvious choice. Consider having the 'original' joint model which is fit to the set of observed data constructed by available measurements for n subjects, $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_n\}$ where $\mathcal{D}_i = \{T_i, \Delta_i, \boldsymbol{Y}_i\}$. Subsequently, we sample with replacement across the *subjects*, such that for each bootstrap replicate $b = 1, \ldots, N$ we obtain a set of 'new' data denoted $\mathcal{D}^{(b)}$. It is important to em-

phasise that the sampling is done over the subjects, and not individual data points which one may carry out in the case of ungrouped data. For each set of 'new' data, we obtain $\hat{\boldsymbol{\Omega}}^{(b)}$ by fitting a joint model to said data. Sample quantities of interest (median value, standard deviation, 95% CI etc.) can then be trivially found by calculation over the N sets of bootstrapped MLEs.

Despite ease of carrying out the bootstrap method, there are several reasons why this avenue is unattractive. Firstly, bootstrapping will obviously be a slow method, and N typically needs to be quite large, with its computation time dependent on the time taken for the EM algorithm itself. However Hickey et al. (2018a) point out simply starting the EM algorithm on resampled data at the the MLEs obtained by the joint model fit to the original data lessens this burden. Secondly, there are no guarantees that resampling won't produce a problematic set of data which could lead to convergence problems in the EM algorithm. Finally both Hsieh et al. (2006) and Xu et al. (2014) note that the bootstrap method overestimated the standard errors in the joint model.

2.4.2 Observed empirical information matrix

Hickey et al. (2018a) approximate the so-called observed empirical information matrix $\mathcal{I}_e$ to obtain SEs for $\hat{\boldsymbol{\Omega}}$. After convergence of the EM algorithm, we have the vector of parameter estimates *and* the estimates for the unspecified baseline hazard $\hat{\lambda}_0(t)$. Using the Breslow estimator for the baseline hazard given in (2.11), we calculate

$$\mathcal{I}_e\big(\hat{\boldsymbol{\Omega}}\big) = \left\{ \sum_{i=1}^{n} s_i\big(\hat{\boldsymbol{\Omega}}\big) s_i^{\top}\big(\hat{\boldsymbol{\Omega}}\big) \right\} - n^{-1} S\big(\hat{\boldsymbol{\Omega}}\big) S^{\top}\big(\hat{\boldsymbol{\Omega}}\big), \tag{2.29}$$

where $s_i\big(\hat{\boldsymbol{\Omega}}\big)$ denotes the gradient vector of the conditional expectation of complete data (profile) log-likelihood function (2.10) evaluated at $\hat{\boldsymbol{\Omega}}$ i.e. the score statistic and $S\big(\hat{\boldsymbol{\Omega}}\big) = \sum_i s_i\big(\hat{\boldsymbol{\Omega}}\big)$. We note that whilst the bootstrap surveyed in the previous section is known to *overestimate* the standard errors, those produced by the approximation of the information matrix here are known to *underestimate* the uncertainty in our parameter estimates owing to the usage of the profile log-likelihood outlined previously in the obtention of score vectors (Hsieh et al., 2006).

2.4.3 Numerical differentiation of the Fisher score

Xu et al. (2014), like the method outlined in the previous section, seek to approximate $\mathcal{I}_o$, and provide four methods based on numerical differentiation to do so. All four presented use either the forward differencing method, or the Richardson extrapolation. Said

differentiation is carried out on either the M-step update (2.9) or the Fisher score vector of the complete data log-likelihood; in both cases a profile technique is used. Owing to its similarity with the method shown in Section 2.4.2, we present only the approach using the profile Fisher score vector and present numerical differentiation techniques in Appendix A.1.

Essentially, Xu et al. (2014) recommend that for each element of $\hat{\boldsymbol{\Omega}}$ which undergoes perturbation in the numerical differentiation routine of choice, the baseline hazard is then re-maximised at this perturbed vector, and the differentiation of this profile Fisher score carried out as normal. Therefore, computational expense is related to the number of parameters and/or the number of random effects as well as the choice of differentiation routine. After convergence, we have the vector of MLEs $\hat{\boldsymbol{\Omega}}$ which has length

$$L = \text{len}\big(\hat{\boldsymbol{\Omega}}\big) = \frac{q(q+1)}{2} + p_1 + \cdots + p_K + p_s + K, \tag{2.30}$$

determined by the joint model specification. In order to obtain the standard errors of $\hat{\boldsymbol{\Omega}}$ by the method in Xu et al. (2014) we perturb, in turn, each of the $l = 1, \ldots, L$ parameters under the numeric differentiation routine. For each perturbation per parameter, we re-calculate the baseline hazard $\hat{\lambda}_0^{(l)}$, and evaluate the profile Fisher score at this l^{th} perturbed vector, thereby obtaining the l^{th} column in the approximated $\mathcal{I}_o$, which is then formed 'over' the numerical differentiation routine. This method of obtaining SEs is notably less computationally expensive under forward differencing, since only one perturbation occurs, compared to the four undertaken by the Richardson extrapolation.

2.5 Simulation

Simulation studies are important statistical tools: Investigating performance properties and adequacy of statistical models in pre-specified situations. We briefly illustrate how we undertake simulation for the longitudinal and survival sub-models before outlining how we simulate data under the 'combined' joint modelling framework. The method(s) outlined are then carried forward in all simulated datasets used. In practice we employ bespoke code which allows for fine-tuning of simulation scenarios we present.

The remainder of this section briefly outlines data simulation under a linear mixed model; simulation of event times; and simulation under the joint modelling framework. We bring the chapter to a close by outlining numerous considerations one may take into account when simulating the data under a joint model.

2.5.1 Simulating the longitudinal response

Simulation of a response under a linear mixed effects model (2.1) is relatively straightforward. For some candidate parameter vector $\boldsymbol{\Omega}^{(\text{TRUE})}_{\text{D},\boldsymbol{\beta},\sigma^2_\varepsilon} = \left(\text{vech}(\text{D})^\top, \boldsymbol{\beta}^\top, \sigma^2_\varepsilon\right)^\top$, which denotes the 'true' values of the necessary parameters, we carry out the following steps:

1. Simulate the random effects $\boldsymbol{b}_i$ for all n desired subjects. In practise, the function `mvrnorm` from the `R` package `MASS` (Venables and Ripley, 2002) is used with zero-mean and the positive semi-definite covariance matrix D supplied.

2. Define the design matrices X_i and Z_i. For all simulations, X_i is defined – unless otherwise stated – as the matrix formed by the horizontal concatenation of an intercept, the vector of follow-up times, a single standard normal deviate and a single Bernoulli draw ($p = 0.5$). We define the vector of follow-up times for each subject as $\boldsymbol{t}_i = (0, \ldots, \kappa)$ with κ the maximal follow-up time. The vector t_i is regularly spaced according to the length of the profile r. Both κ and r are controlled in simulations. The specification of Z_i is simulation dependent: For instance if a random intercept and slope random is desired, then it is defined as the horizontal concatenation of the intercept column and $\boldsymbol{t}_i$.

3. Define $\boldsymbol{\eta}_i = \text{X}_i\boldsymbol{\beta} + \text{Z}_i\boldsymbol{b}_i$ and simulate $\boldsymbol{Y}_i$ using `rnorm` with mean $\boldsymbol{\eta}_i$ and standard deviation $\sqrt{\sigma^2_\varepsilon}$.

The above steps then produce n sets of r-vectors of the simulated Gaussian response. That is, $\mathbb{E}_i\left[\boldsymbol{Y}_i | \boldsymbol{b}_i; \boldsymbol{\Omega}^{(\text{TRUE})}_{\text{D},\boldsymbol{\beta},\sigma^2_\varepsilon}\right]$ is conditionally (multivariate) normal. One can then trivially extend this to produce K vectors with differing e.g. fixed effects. In circumstances where the response one wishes to simulate is non-Gaussian, then some requisite inverse-link function is used on $\boldsymbol{\eta}_i$, as we go on to detail in Chapter 4.

Of course, these r-vectors of response(s) are truncated at the subject's simulated event time in the joint model setting. We outline methodology to simulate failure times in the general case in the next section and under a joint model in Section 2.5.3.

2.5.2 Simulation of survival times

Cox proportional hazards regression is the ubiquitous approach to modelling the effect of a the covariate vector $\boldsymbol{S}_i$ on the hazard of some time-to-event outcome. Simplifying the Cox PH regression model (2.2) to the usual $\lambda_i(t|\boldsymbol{S}_i) = \lambda_0(t)\exp\{\boldsymbol{S}_i^\top\boldsymbol{\zeta}\}$, with $\lambda_0(t)$ and $\boldsymbol{\zeta}$ previously defined in Section 2.1.1. The survival function, which gives probability that subject i survives past time t is then $S(t) = \exp\left[-\Lambda_0(t)\exp\{\boldsymbol{S}_i^\top\boldsymbol{\zeta}\}\right]$, where $\Lambda_0 = \int_0^t \lambda_0(u)du$ denotes the cumulative baseline hazard. The distribution of survival times under this Cox

Distribution	Parameter(s)	Formula for simulation
Exponential	Scale $\nu > 0$	$T_i^* = -\frac{\log U}{\nu \exp\{\boldsymbol{S}_i^\top \boldsymbol{\zeta}\}}$
Weibull	Scale $\nu > 0$, shape $\alpha > 0$	$T_i^* = \left(-\frac{\log U}{\nu \exp\{\boldsymbol{S}_i^\top \boldsymbol{\zeta}\}}\right)^{1/\alpha}$
Gompertz	Scale $\nu > 0$, shape $-\infty < \alpha < \infty$	$T_i^* = \frac{1}{\alpha} \log\left[1 - \frac{\alpha \log(U)}{\nu \exp\{\boldsymbol{S}_i^\top \boldsymbol{\zeta}\}}\right]$

Table 2.1: Simulation of event times for a given subject i under the exponential, Weibull and Gompertz distributions. U denotes a uniform draw in each case and $\boldsymbol{S}_i$ is subject i's vector of baseline covariates.

PH regression is then $F(t|\boldsymbol{S}_i) = 1 - S(t|\boldsymbol{S}_i)$. Bender et al. (2005) demonstrate that one can simulate the true event time of subject i by

$$T_i^* = \Lambda_0^{-1}\left[-\log U \exp\left\{-\boldsymbol{S}_i^\top \boldsymbol{\zeta}\right\}\right], \quad U \sim \text{Unif}(0,1)\,.$$

Therefore, survival times can be generated for any baseline hazard which has an invertible cumulative baseline hazard function Λ: Bender et al. (2005) noted that only the exponential, Weibull and Gompertz baseline hazard functions satisfy not only this criterion, but additionally share the assumption of proportional hazards with the Cox PH model. The formula required for simulation of survival times for each of these three candidate distributions are given in Table 2.1

Going forward, owing to its wide use in survival analysis, we use the Gompertz as our baseline hazard of choice. The simulated event time T_i^* is given by

$$T_i^* = \frac{1}{\alpha} \log\left[1 - \frac{\alpha \log(U)}{\nu \exp\{\boldsymbol{S}_i^\top \boldsymbol{\zeta}\}}\right], \quad U \sim \text{Unif}(0,1)\,, \tag{2.31}$$

where α and ν are the shape and scale, respectively, of the Gompertz distribution. As was the case for simulation of the longitudinal process in Section 2.5.1, we are able to fine-tune the survival process by choice of parameter $\boldsymbol{\zeta}$, in addition to the underlying baseline hazard by ν and α, which are defined for the Gompertz distribution in Table 2.1. This fine-tuning is done *independently* of the longitudinal process at this stage; we consider simulation under the joint modelling framework proper in the next section.

We note in passing that another way to simulate T_i^* is to directly solve the survival function $S(t)$ for each subject given their covariates $\boldsymbol{S}_i$ via a root-finding algorithm (such as `uniroot` in base `R`); a useful alternative if the distribution of interest is non-invertible or the baseline hazard is otherwise specified.

2.5.3 Simulation under a joint modelling framework

Advantages of using Cox PH regression are its ability to incorporate both (a): Time-varying covariate effects, that is, a covariate whose value is fixed at baseline, but effect on the hazard is time-varying; and (b): Time-varying covariates, that is a covariate whose value changes over time, e.g. cumulative dosage received in a clinical trial (Austin, 2012). Under the Cox PH specification in the joint modelling framework with a random intercept-and-slope specification we indeed have a time varying covariate in the form of the K slope terms across random effects $\boldsymbol{b}_{i1} = (b_{i11}, b_{i21}, \ldots, b_{iK1})^\top$ thereby falling into category (b). When a random intercepts-only structure is instead desired, the formulation (2.31) suffices, with the term including covariate information expanded to include our association parameters and random intercept terms: $\exp\left\{\boldsymbol{S}_i^\top \boldsymbol{\zeta} + \sum_{k=1}^{K} \gamma_k b_{ik0}\right\}$.

Given inclusion of random slope terms $\boldsymbol{b}_{i1}$, we follow closely the work of Austin (2012), who expand work outlined in the previous section by Bender et al. (2005) to the case with such a time-varying covariate. Rewriting the Cox PH formulation (2.2)

$$\begin{aligned} \lambda(t|\boldsymbol{S}_i, \boldsymbol{b}_i; \boldsymbol{\zeta}, \boldsymbol{\gamma}) =& \lambda_0(t) \exp\left\{\left(\boldsymbol{S}_i^\top \boldsymbol{\zeta} + \sum_{k=1}^{K} \gamma_k b_{ik0}\right) + \left(\sum_{k=1}^{K} \gamma_k b_{ik1}\right) t\right\}, \\ \lambda(t|\cdot) =& \lambda_0(t) \exp\{P + Qt\} \end{aligned} \tag{2.32}$$

where we have collected time-invariant and time-varying terms together into the arbitrarily-named P and Q, respectively. Then, given that the Gompertz baseline hazard is $\lambda_0(t) = \nu \exp\{\alpha t\}$, with shape $-\infty < \alpha < \infty$ and scale parameter $\nu > 0$. We obtain from (2.32),

$$\begin{aligned} \Lambda(t|\cdot) &= \int_0^t \lambda(u|\cdot) du = \int_0^t \exp\{P + Qu\}\, \nu \exp\{\alpha u\}\, du \\ &= \nu \exp\{P\} \int_0^t \exp\{(Q + \alpha)u\}\, du \\ &= \nu \exp\{P\} \left[\frac{1}{Q + \alpha} \exp\{(Q + \alpha)\, u\}\right]_0^t \\ \Lambda(t|\cdot) &= \frac{\nu \exp\{P\}}{Q + \alpha} [\exp\{(Q + \alpha)\, t\} - 1]\,. \end{aligned}$$

Then, setting the survival function $S(t)$ equal to a random uniform draw, U, and rear-

ranging for t we obtain

$$U = S(t) = \exp\left\{-\frac{\nu \exp\{P\}}{Q+\alpha}[\exp\{(Q+\alpha)\,t\} - 1]\right\}$$
$$\exp\{(Q+\alpha)\,t\} = 1 - \frac{\nu \exp\{P\}}{Q+\alpha}\log U$$
$$\implies t = \log\left(1 - \frac{Q+\alpha}{\nu \exp\{P\}}\log U\right) \Big/ (Q+\alpha)\,.$$

Finally then, our true event time T_i^* can be generated by

$$T_i^* = \frac{1}{Q+\alpha}\log\left[1 - \frac{(Q+\alpha)\log U}{\nu \exp\{P\}}\right], \quad U \sim \text{Unif}(0,1)\,, \tag{2.33}$$

where P and Q are defined in (2.32). Similar derivations for Weibull and exponential baseline hazards are given in Appendix A.2.

In addition to the survival time as outlined in this section and the last, we additionally generate n censor-times, $C_i \sim \text{Exp}(\Upsilon)$ where Υ denotes the rate parameter; representing yet more control over the simulated data. With both what is the 'true' failure time T_i^* and censor time C_i simulated, we set each observed event time $T_i = \min(T_i^*, C_i)$ and generate Δ_i accordingly. If $T_i > \kappa$ then it is truncated to value $T_i = \kappa + 0.1$ for analysis purposes and Δ_i is set to zero. The baseline covariates used in simulating the failure time for subject i throughout the thesis are the same single standard normal and Bernoulli draws obtained in Section 2.5.1.

2.5.4 Simulation beyond an intercept and slope structure

The mechanisms for simulation for a joint model under an intercept-only and intercept-and-slope random effects structure has been outlined in the previous section: The intercept-only random effects formulation is encapsulated in previous work by Bender et al. (2005) presented in (2.31), and the inclusion of a random slope in addition lead us to consider work by Austin (2012) in treating these as time-varying covariates (2.33).

However, simulations are also necessary when working with more complex random effects structures. For instance, under a quadratic formulation in the simulated response, or when spline terms are included with M internal knots, taking form $\sum_{m=0}^{M}(\alpha_m + b_{im})B_m(t)$, where $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_M)$ are B-spline coefficients and $\{B(t)\}$ the truncated linear basis (see e.g. Barrett and Su (2017) for a joint model with such a spline specification). Under such examples, the integration outlined in the previous section becomes awkward and unappealing, and so we utilise previously-used methods (Philipson et al., 2018) to generate data under more complex random effects structures.

We expand the time-varying term $Q \to Q(\boldsymbol{b}_i, t)$ here to encapsulate any functional form of the random effects. For example, utilising a quadratic random effects structure arbitrarily across all k responses results in

$$Q(\boldsymbol{b}_i, t) = \sum_{k=1}^{K} \gamma_k \left(b_{ik1} t + b_{ik2} t^2 \right),$$

and likewise any polynomial of order p gives

$$Q(\boldsymbol{b}_i, t) = \sum_{k=1}^{K} \gamma_k (b_{ik1} t + \cdots + b_{ikp} t^p)$$

as well as e.g. spline term(s). Given the definition of the hazard $\lambda(t|\cdot)$ is

$$\lim_{\Delta t \to 0} \frac{\Pr(t \leq T \leq t + \Delta t | T \geq t, \ \cdot)}{\Delta t}, \tag{2.34}$$

which we can trivially rearrange to

$$\lim_{\Delta t \to 0} \lambda(t|\cdot) \Delta t = \lim_{\Delta t \to 0} \Pr(t \leq T \leq t + \Delta t | T \geq t, \ \cdot), \tag{2.35}$$

with the hazard $\lambda(t|\cdot)$ given by (2.32). Therefore failure times, simulated under a host of random effects specifications given by $Q(\boldsymbol{b}_i, t)$, are available.

Choosing a nominally small value for Δt allows us to generate a sufficiently fine grid which multiplies the hazard. This scaled hazard is equivalent to the probability of failure occurring in the fine-grid time interval $(t, t + \Delta t)$. We also generate a vector of possible times for each individual from time 0 to the truncation time κ with step-size Δt. Each of the generated probabilities can be compared to a Unif$(0, 1)$ draw. If the probability does not exceed the uniform deviate then the survival time is set as κ, otherwise the survival time takes the value of the minimum candidate time (i.e. when the event first happened).

To illustrate, consider subject i with step-size $\Delta t = 0.02$ and truncation time $\kappa = 5$. We would generate vector of candidate times $\boldsymbol{t} = (0.00, 0.02, \ldots, 2.50, 2.52, \ldots, 4.98, 5.00)^\top$. The probability of an event occurring at each $t \in \boldsymbol{t}$ is then calculated, dependent on form of chosen $Q(\boldsymbol{b}_i, t)$. These probabilities are then compared against Unif$(0, 1)$ realizations to obtain, say, $\boldsymbol{t}^* = (\kappa, \kappa, \ldots, \kappa, 2.52, \ldots, 4.98, \kappa)^\top$. The minimal value of $\boldsymbol{t}^*$ is then set as the true failure time for this subject, $T_i^* = 2.52$. Subject i additionally has censor-time simulated as outlined in Section 2.5.3, which allows us to allocate T_i in an identical manner.

Although this methodology allows for a wide breadth of random effects structures, the use of potentially very large (owing to elected step-size) matrices can slow down data gener-

ation considerably. Additionally, the vector of candidate times $\boldsymbol{t}$ are effectively discrete times, which could have the effect of many subjects having the same simulated failure time, which could be troublesome in certain simulation scenarios. Of course, this could be circumvented by choosing a finer value for Δt, at greater computational cost.

2.6 Simulation considerations

Simulation of survival times is a task which is hard to control with absolute perfection: More so an act of balancing several different aspects which come from the specification of parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\zeta}$; the underlying Gompertz baseline hazard via scale and shape parameters ν and α, respectively; as well as the magnitude of the random effects via their allocated variance in D. A delineation of how these different facets – which we control in simulation settings – combine to allow one to obtain n failure times time was given in Section 2.5.3. In this section we consider these three factors which can alter both the incidence of failure in $[0, \kappa]$ as well as the distribution of simulated T_i^*.

2.6.1 The baseline hazard

We exclusively simulate under a Gompertz baseline hazard of the form $\lambda_0(t) = \nu \exp\{\alpha t\}$. The scale parameter ν can be thought of as baseline hazard's intercept term: If the shape parameter α is zero then the baseline hazard is stationary through time, and values of γ and ζ mainly drive the failure times with no inclusion of time itself contributing to the underlying baseline hazard. The shape parameter itself therefore allows the baseline hazard to increase or decrease over time.

Setting the survival parameters to zero, such that they do not impact the hazard, we select a set of values for $\log \nu = \{-4, -3, -2, -1\}$ and $\alpha = \{-0.4, -0.2, 0.2, 0.4\}$ and simulate one hundred sets of data, collating the simulated failure times from each simulated data set.

Figure 2.1 shows the distribution of survival times under each combination of $\{\log \nu\}$ and $\{\alpha\}$ above. At lower values of $\log \nu$ we note the distribution of event times changes as α increases from a negative to positive value: Failures occur earlier on in follow-up (driven by $\log \nu$) with visible drop-off in frequency over time owing to the negative α, and vice versa for $\alpha > 0$.

For $\log \nu >= -2$ this change in distribution is less pronounced owing to the underlying Gompertz baseline hazard (arising from this scale parameter) being already adequately high. The failure times are more concentrated around the start of simulated follow-up at largest $\alpha = 0.4$, whereas for $\alpha = -0.4$ there is still a somewhat noticeable 'tail' of failure times comparatively.

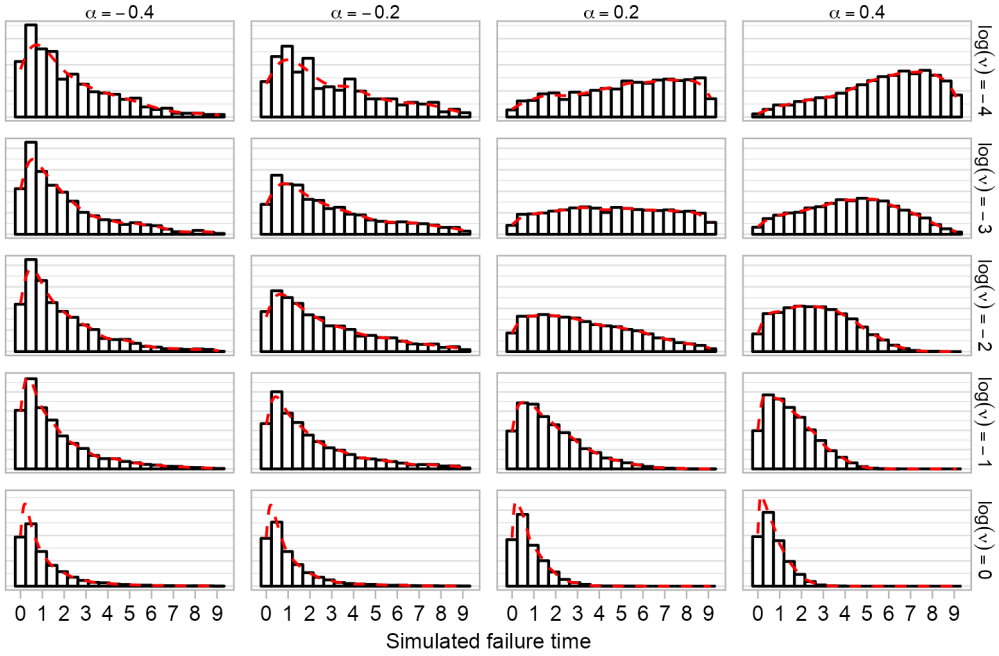


Figure 2.1: Distribution of simulated survival times under different parameterisations of the Gompertz baseline hazard.

2.6.2 Magnitude of survival parameters

Next, we consider altering the magnitude and sign of the survival parameters. Specifically, we simulate a univariate model with candidate association parameters $\gamma = \{-1.0, -0.5, 0.0, 0.5, 1.0\}$ and the coefficient attached to the binary baseline covariate in the survival model, taking values $\zeta_2 = \{-1.0, -0.3, 0.3, 1.0\}$. Across each of these combinations of $\{\gamma\}$ and $\{\zeta_2\}$, we hold the parameters controlling the baseline hazard constant at $\log \nu = -1$ and $\alpha = 0$ in order to simulate a time-invariant baseline hazard. The covariance matrix on the random effects is set as a diagonal matrix of dimension two. With these set out, we expect the changes in simulated failure time densities to be driven by the changing values of $\{\gamma\}$ and $\{\zeta_2\}$.

Figure 2.2 illustrates the effect of different magnitudes of γ and ζ_2. The effect of an increased ζ_2 is perhaps best exemplified in the 'column' for $\gamma = 0$. Here, we infer that the increasing value of ζ_2 simply increases the underlying baseline hazard, illustrated by the greater incidence of failures due to said higher baseline hazard. Interestingly, the distribution of failure times for the positive-negative pairs of values for γ are visually (almost) identical. This is due to the simulation of the random effects producing an approximate half-and-half 'split' of those who are above the global trajectory (thus their hazard *increases* with positive γ) and those below it (*increasing* with negative γ). We see

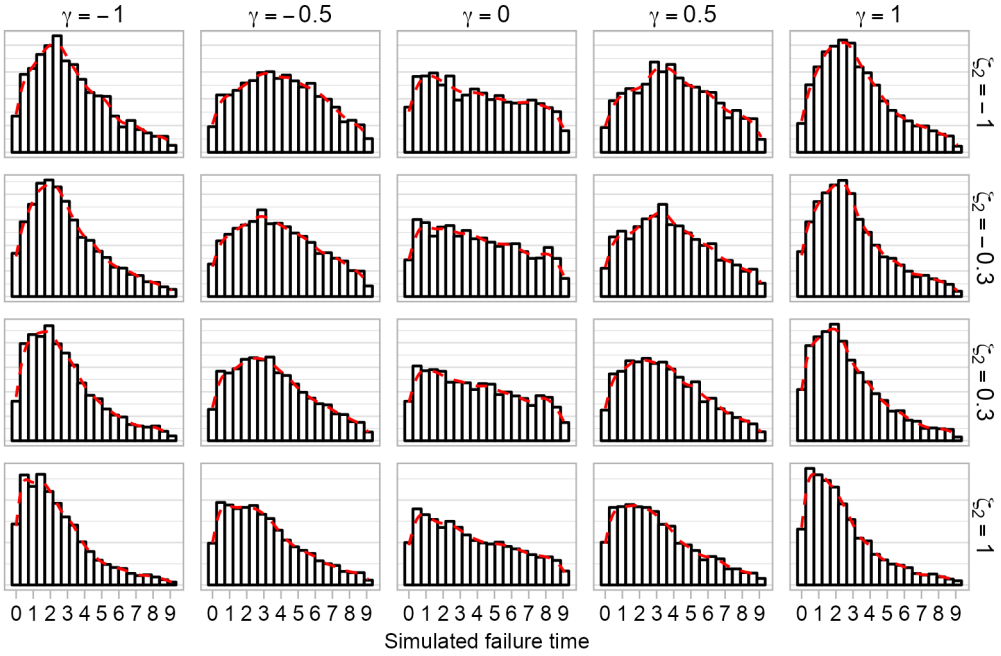


Figure 2.2: Distribution of simulated survival times under different parameterisations of the survival parameters γ and ζ_2.

as $|\gamma|$ increases, more failures occur earlier in the period of follow-up, which is exacerbated by higher values of ζ_2.

2.6.3 Magnitude of the random effects

In a similar spirit to our investigation of the effect of survival parameter magnitude, we set the baseline hazard constant at $\log \nu = -1$ and $\alpha = 0$. We hold the association parameter constant at a moderate $\gamma = 0.5$ and $\boldsymbol{\zeta} = \mathbf{0}$, such that failures occurring are likely to be due to ones deviation from a global trend, as dictated by the parameterisation of the covariance matrix D. We set off-diagonal parameters $\mathrm{D}_{12} = \mathrm{D}_{21} = 0$ and consider all 16 possibilities for the diagonal of D from candidate intercept variances $\mathrm{D}_{11} = \{3.0, 1.0, 0.5, 0.5^2\}$ and slope variances $\mathrm{D}_{22} = \{1.0, 0.5, 0.25, 0.25^2\}$.

Figure 2.3 illustrates how the distribution of simulated survival times changes when the variance of random effects (thus the average magnitude of the random effects) increases. When the random effects can take larger values (i.e. D_{11} and/or D_{22} are larger), an overall increase in the number of failure times as time progresses occurs. When the random intercept variance is greatest at $D_{11} = 3$ the simulated failure times are most concentrated at start of follow-up occurs, especially when D_{22} is comparatively smaller. Conversely, when random slope variance is greater, the distribution of simulated failure times have an

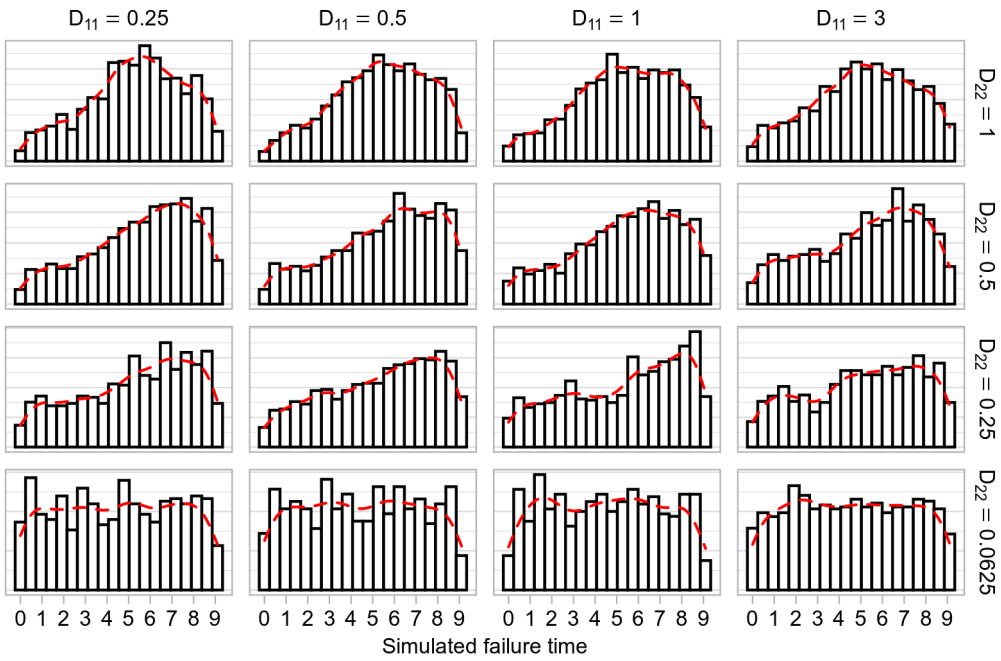


Figure 2.3: Distribution of simulated survival times under different magnitudes of random effect variance.

elongated tail, demonstrating the impact of the simulated random slopes on the hazard.

Chapter 3

Faster Fitting: An Approximate EM Algorithm

The work presented in this chapter is based on the publication:
Murray, J., Philipson, P., 2022. *A fast approximate EM algorithm for joint models of survival and multivariate longitudinal data.* Computational Statistics & Data Analysis 170, 107438. doi: 10.1016/j.csda.2022.107438.

3.1 Motivation

In Chapter 2 we outlined the multivariate joint model framework with K longitudinal responses which are assumed continuous and Gaussian. Available software for such joint models, representing existing computational approaches, was outlined in Sections 1.2.3 and 1.2.4 encompassed both maximum likelihood approaches as well as the Bayesian paradigm.

Multivariate joint models (MVJMs) are superior to K separate univariate fits, as they take into account the correlation *between* the longitudinal responses; using all relevant information in the joint model thus obtaining correctly adjusted estimates for each response. Additionally, one obtains a single prediction using all possible information, rather than several from said K univariate fits, each of which is likely to overstate the importance of any association when treated in isolation.

Fitting of these models is computationally burdensome, as mentioned in Section 2.3.1, due to the multidimensional integral present in the E-step (2.13).

3.1.1 E-step constraints in multivariate joint models

As we outlined in Section 2.3, and as alluded to above, most computational issues arise due the E-step (2.10). Here necessary expectations are taken with respect to the distribution of the random effects conditional on the observed data at the current set of parameter estimates. Specifically, we require expectations of the form (2.13), with this conditional distribution then having form (2.16). However, owing to either an increased number of longitudinal responses in the multivariate joint model of interest; the complexity of their random effects structures; or both, existing techniques are likely to be less viable due to the inherent computational burden. This is to say, the existing methods we outlined in Sections 2.3.2–2.3.5 are likely to become infeasible as the dimension of the random effects, q, increases. Said intractability often manifesting itself by the computation time taken to fit these multivariate joint models being commensurate with q: Limiting the application of multivariate joint models to smaller numbers of random effects, longitudinal processes, or both.

The different methods we outlined to evaluate integrals of interest in Sections 2.3.3–2.3.5 in fact were all proposed in turn to overcome the apparent computational burden in earlier literature e.g. Wulfsohn and Tsiatis (1997). Adaptive quadrature methods sought to reduce the number of weights and abscissae used in numerical integration, but the amount of evaluations may still be computationally demanding in spite of this, owing to large q. Monte Carlo methods themselves do not suffer the same burden, but may become prohibitive for higher dimensional random effects owing to the number of required samples

for adequate evaluation.

Hickey et al. (2018a) go as far to speculate that inference for joint models with very many longitudinal responses, such as electronic healthcare records, may require some approximate method in evaluation of the integral (2.13) as part of parameter estimation under maximum likelihood.

3.2 An approximate EM algorithm

3.2.1 A normal approximation

Bernhardt et al. (2015) proposed a normal approximation on the distribution of the random effects $f(\boldsymbol{b}_i|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega})$ for each individual conditional on the observed data in the context of a joint model with a *logistic* sub-model in place of the survival one (2.2). The proposed normal approximation has the effect of reducing the dimensionality of each required integral to be uniformally one, regardless of the complexity of the random effects. This then exploits that any linear combination of $\boldsymbol{b}_i$ would *also* be normal, thereby improving computational efficiency. We seek then to extend their approximation to the joint model with a survival sub-model as introduced in the previous Chapter in Sections 2.1.1 and 2.1.2.

Bernhardt et al. (2015) propose, at the parameter estimates at the current iteration (m), $\boldsymbol{\Omega}^{(m)}$,

$$\boldsymbol{b}_i|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega}^{(m)}, \lambda_0^{(m)} \overset{\text{appx.}}{\sim} N\big(\hat{\boldsymbol{b}}_i, \hat{\Sigma}_i\big), \tag{3.1}$$

where $\hat{\boldsymbol{b}}_i$ is the vector which maximises the complete data log-likelihood at the current set of parameter estimates:

$$\hat{\boldsymbol{b}}_i = \underset{\boldsymbol{b}_i}{\operatorname{argmax}} \log f\big(\boldsymbol{b}_i, T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega}^{(m)}\big); \tag{3.2}$$

with estimated variance

$$\hat{\Sigma}_i = \left\{ -\frac{\partial^2 \log f\big(\boldsymbol{b}_i, T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega}^{(m)}\big)}{\partial \boldsymbol{b}_i \partial \boldsymbol{b}_i^\top} \bigg|_{\boldsymbol{b}_i = \hat{\boldsymbol{b}}_i} \right\}^{-1}. \tag{3.3}$$

In (3.1) we explicitly condition on the current estimate for the baseline hazard, $\lambda_0^{(m)}$, which we treat as a nuisance parameter, holding *implicit* membership in $\boldsymbol{\Omega}^{(m)}$ only. Hereafter, its involvement in the approximation and parameter estimates continues to be implicit in nature.

It was previously demonstrated by Rizopoulos (2012a) that $f(\boldsymbol{b}_i|\boldsymbol{Y}_i;\boldsymbol{\Omega})$ is approximately normal as $m_i \to \infty$ within the confines of a 'classic' *univariate* joint model. Subsequently, Bernhardt et al. (2015) extended this to the multivariate Gaussian case we considered in Chapter 2 for the 'full' conditional distribution shown in (3.1). We later devote Chapter 5 to the justification of the above approximation.

The approximation (3.1) then allows for *all* necessary conditional expectations which are of the form $\mathbb{E}_i\Big[g(\boldsymbol{b}_i)\,|T_i,\Delta_i,\boldsymbol{Y}_i;\boldsymbol{\Omega}^{(m)}\Big]$ to be taken with respect to a univariate normal distribution. Said normal distributions can then be quickly evaluated using adaptive Gauss-Hermite quadrature, which we outline in Section 3.3.

3.2.2 Starting values

The number of iterations required by the EM algorithm can be reduced by choosing starting values, $\boldsymbol{\Omega}^{(0)}$, which are close to the maximiser. We set about obtaining these initial values using pre-existing `R` packages.

For $k = 1,\ldots,K$ we obtain parameter estimates for the k^{th} longitudinal process by fitting a linear mixed effects model using `glmmTMB` (Brooks et al., 2017). This returns initial conditions for the fixed effects $\boldsymbol{\beta}_k^{(0)}$ and the residual variance $\sigma_{\varepsilon_k}^{2^{(0)}}$. Additionally, we obtain the best linear unbiased predictors (BLUPs) for the random effects $\boldsymbol{b}_{ik}$ along with their variance-covariance matrix $\mathrm{D}_k^{(0)}$. Then, across all K responses we can construct the initial condition for the block-diagonal $\mathrm{D}^{(0)} = \bigoplus_{k=1}^{K}\mathrm{D}_k^{(0)}$, and populate the off-block-diagonal elements with $\operatorname{cov}(\boldsymbol{b}_e,\boldsymbol{b}_f)\,,\ e,f = 1,\ldots,K,\ e \neq f$.

Next, we use the aforementioned BLUPs for each $k = 1,\ldots,K$ response, $\boldsymbol{b}_{ik}$, as K independent time-varying covariates in the usual Cox PH model, along with the time invariant $\boldsymbol{S}_i,\ i = 1\ldots n$, to obtain $\boldsymbol{\Phi}^{(0)} = \left(\gamma_1^{(0)},\ldots,\gamma_K^{(0)},\boldsymbol{\zeta}^{(0)^\top}\right)^\top$. The form of each of these time-varying covariates is dependent upon the form of $\boldsymbol{W}_k(\cdot)$ in (2.2) (i.e. intercept and slope, quadratic specification and so on). This is done using the `survival` package (Therneau, 2015), and additionally returns the initial estimate for the baseline hazard, $\lambda_0^{(0)}(\cdot)$.

We can then form $\boldsymbol{\Omega}^{(0)}$ and go about parameter estimation via the EM algorithm, confident that it commences at parameter values which are conducive to faster convergence of the algorithm itself. We used `glmmTMB` as it allows for efficient generalised mixed model fitting for numerous families, and we found it to provide much more stable parameter estimates in a more timely manner when compared with competing packages. Other packages, such as `nlme` (Pinheiro et al., 2021) or `lme4` (Bates et al., 2015), are very slightly quicker but restricted to modelling continuous (Gaussian) $\boldsymbol{Y}_i$, since we progress to considering non-Gaussian responses in Chapter 4, we additionally use `glmmTMB` here for simplicity.

3.2.3 Convergence details

Posit that we have just completed one EM step from iteration (m) to iteration $(m+1)$, and want to ascertain whether the algorithm has converged. Two standard rules for such an investigation are the absolute and relative convergence criteria – the existing packages `joineR` (Philipson et al., 2018) using the former and `JM` (Rizopoulos, 2010) the latter – for instance.

An issue one may encounter when attempting to maximise the likelihood via an iterative procedure such as EM is in becoming 'stuck' near the maximiser, or prematurely declaring convergence to it. This can be remedied by taking into account the scale of parameters when applying these stopping rules: If a parameter is very small (i.e. close to zero) in magnitude, it is perhaps likely that some numerical issue may come into play in calculation of the relative difference; likewise the absolute difference doesn't take into account the scale of parameters at all, and may prematurely stop the algorithm.

One method to circumvent these potential issues is to employ a two-pronged approach:

$$\begin{cases} \max\left(|\Omega_1^{(m+1)} - \Omega_1^{(m)}|, \ldots, |\Omega_L^{(m+1)} - \Omega_L^{(m)}|\right) & < \xi_1 \text{ if } |\Omega_x| < \upsilon \\ \max\left(\frac{|\Omega_1^{(m+1)} - \Omega_1^{(m)}|}{|\Omega_1^{(m)}| + \nu}, \ldots, \frac{|\Omega_L^{(m+1)} - \Omega_L^{(m)}|}{|\Omega_L^{(m)}| + \nu}\right) & < \xi_2 \text{ if } \underset{x=1,\ldots,L}{|\Omega_x| \geq \upsilon}. \end{cases} \tag{3.4}$$

where L denotes the dimension of $\boldsymbol{\Omega}$ (2.30). Interestingly, this is the stopping rule used by the software `SAS` in its EM algorithms; repurposed in the joint modelling setting recently by `joineRML` (Hickey et al., 2018a).

We opt for the convergence criterion shown above as the approach precludes issues one may encounter when considering solely one convergence criterion, whilst still affording accuracy in those parameters with estimates closer to zero; in particular those variance components on the diagonal of D that are bounded below by zero. In the relative difference criterion, ν is some small value added to the denominator to preclude numerical issues which might occur in the calculation here. In all presented simulations and analyses we use $\xi_1 = \nu = 10^{-3}$, $\xi_2 = 5 \times 10^{-3}$ and $\upsilon = 0.1$ unless otherwise stated.

3.2.4 The approximate EM algorithm

With details on how we obtain starting values and determine whether or not the model has converged, we now set out to illustrate how we utilise the approximation in Section 3.2.1 within the context of an EM algorithm for a joint model.

In short, we obtain our initial conditions for the parameter vector $\boldsymbol{\Omega}$ and iteratively cycle between the E- and M-steps, refining these parameter estimates until convergence is

deemed to have occurred. At *each* iteration we utilise the approximation (3.1)–(3.3) to obtain conditional expectations before forming parameter updates; this specific process being elucidated in Section 3.3.

With that being said, we lay out the general steps taken to fit MVJMs outlined in Section 2.1 in a similar spirit to existing work (Bernhardt et al., 2015; Murray and Philipson, 2022, 2023):

1. Employ the methodology outlined in Section 3.2.2 to obtain the initial conditions for the parameters, $\boldsymbol{\Omega}^{(0)}$.
2. For any iteration $(m+1)$:
   i. Maximise $\log f\big(\boldsymbol{b}_i, T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega}^{(m)}\big)$ using the `optim` function with the quasi-Newton BFGS algorithm to obtain $\hat{\boldsymbol{b}}_i$ and $\hat{\Sigma}_i$ (see (3.2) and (3.3), respectively).
   ii. Use the approximation (3.1) to update the parameter vector $\boldsymbol{\Omega}^{(m)} \rightarrow \boldsymbol{\Omega}^{(m+1)}$; greater detail is provided in Section 3.3.
3. Check whether the algorithm has converged by evaluating the convergence criterion (3.4). If the criterion is satisfied, exit the algorithm; otherwise, proceed to the next iteration by setting $(m) \rightarrow (m+1)$.
4. Repeat steps 2. and 3. for a minimum of four iterations, cycling between parameter updates and convergence checks.

3.2.5 Standard error calculation

In Section 2.4 we explored three methods to calculate standard errors in order to complete inference. There, we posited that the approximation of the observed empirical information matrix (2.29) would likely be preferable from a computational standpoint: It offers a route to obtain the standard errors in a more timely manner than by bootstrapping or repeated numerical differentiation.

The observed empirical information matrix (2.29) is calculated by n subject specific score vectors evaluated at the MLEs $\hat{\boldsymbol{\Omega}}$

$$s_i\big(\hat{\boldsymbol{\Omega}}\big) = \int_{-\infty}^{\infty} \left[\frac{\partial}{\partial \hat{\boldsymbol{\Omega}}} \Big\{ \log f\big(\boldsymbol{Y}_i | \boldsymbol{b}_i; \hat{\boldsymbol{\Omega}}\big) + \log f\big(T_i, \Delta_i | \boldsymbol{b}_i; \hat{\boldsymbol{\Omega}}\big) + \log f\big(\boldsymbol{b}_i | \hat{\boldsymbol{\Omega}}\big) \Big\} \right] f\big(\boldsymbol{b}_i | T_i, \Delta_i, \boldsymbol{Y}_i; \hat{\boldsymbol{\Omega}}\big) d\boldsymbol{b}_i, \tag{3.5}$$

wherein, given the approximate method outlined in 3.2.1, the expectation presented above is taken with respect to the normal distribution in (3.1).

Specifically, after convergence of the EM algorithm, we find $\hat{\boldsymbol{b}}_i$ and $\hat{\Sigma}_i$ at the purported MLEs, $\hat{\boldsymbol{\Omega}}$. Then, the (profile) score vector (3.5) is calculated. The constituent densities herein are conditionally independent given the random effects $\hat{\boldsymbol{b}}_i$, therefore we form 'individual' constituent scores to form the score vector

$$s_i\big(\hat{\boldsymbol{\Omega}}\big) = \Big(s_i\big(\hat{\boldsymbol{\Omega}}_{\mathrm{D}}\big)^\top, s_i\big(\hat{\boldsymbol{\beta}}\big)^\top, s_i\big(\hat{\sigma}^2_{\varepsilon_1}\big), \ldots, s_i\big(\hat{\sigma}^2_{\varepsilon_K}\big), s_i\big(\hat{\boldsymbol{\Phi}}\big)^\top\Big)^\top,$$

wherein $\hat{\boldsymbol{\Omega}}_{\mathrm{D}} \equiv \mathrm{vech}\big(\hat{\mathrm{D}}\big)$. The constituent scores are given as (for simplicity's sake, the hat notation is momentarily dropped):

$$\begin{aligned}
s_i(\boldsymbol{\Omega}_{\mathrm{D}}) &= -\frac{1}{2}\mathrm{Tr}\left(\mathrm{D}^{-1}\frac{\partial \mathrm{D}}{\partial \boldsymbol{\Omega}_{\mathrm{D}}}\right) + \frac{1}{2}\mathrm{Tr}\left(\mathrm{D}^{-1}\frac{\partial \mathrm{D}}{\partial \boldsymbol{\Omega}_{\mathrm{D}}}\mathrm{D}^{-1}\mathbb{E}_i\left[\boldsymbol{b}_i\boldsymbol{b}_i^\top\right]\right);\\
s_i(\boldsymbol{\beta}) &= \mathrm{X}_i^\top \mathrm{V}_i^{-1}(\boldsymbol{Y}_i - \mathrm{X}_i\boldsymbol{\beta} - \mathrm{Z}_i\mathbb{E}_i[\boldsymbol{b}_i]);\\
s_i\big(\sigma^2_{\varepsilon_k}\big) &= -\frac{m_{ik}}{2\sigma^2_{\varepsilon_k}} + \frac{\mathbb{E}_i\left[(\boldsymbol{Y}_i - \mathrm{X}_i\boldsymbol{\beta} - \mathrm{Z}_i\boldsymbol{b}_i)^\top(\boldsymbol{Y}_i - \mathrm{X}_i\boldsymbol{\beta} - \mathrm{Z}_i\boldsymbol{b}_i)\right]}{2\sigma^4_{\varepsilon_k}};\\
s_i(\boldsymbol{\Phi}) &= \frac{\partial \mathbb{E}_i[\log f(T_i, \Delta_i|\boldsymbol{b}_i;\boldsymbol{\Omega})]}{\partial \boldsymbol{\Phi}};
\end{aligned} \tag{3.6}$$

where $\frac{\partial \mathrm{D}}{\partial \boldsymbol{\Omega}_{\mathrm{D}}}$, the derivative of matrix D with respect to its half-vectorisation, is a matrix itself and represents the location of those elements (i.e. a matrix of zeroes with ones placed according to the element of the half-vectorisation, thereby capturing their contribution to calculation of the score). An alternative form for the score on vech(D) is given in Appendix A.4. Of course, in actuality the expectations in (3.6) are conditioned on the observed data and the maximum likelihood estimates $\hat{\boldsymbol{\Omega}}$ as was the case in e.g. (2.13); the form of such conditional expectations under the normal approximation is elucidated in the next section.

We additionally draw attention to the method proposed by Xu et al. (2014), outlined in Section 2.4.3, being an unattractive route: We systematically perturb each of the L parameters constructing $\boldsymbol{\Omega}$ (2.30) in turn; each requiring maximisation of the baseline hazard, followed by calculation of n sets of $\left\{\hat{\boldsymbol{b}}_i, \hat{\Sigma}_i\right\}$. Therefore, under this proposed method, obtaining the information matrix by forward differencing (Appendix A.1, the 'cheaper' option) requires $n \times (1 + L)$ `optim` calls to find the modal value and its variance alone.

3.3 The M-step

In Section 2.2.2 we outlined the general form of parameter updates which form the M-step (and elucidate the form of requisite conditional expectations required) in (2.11) and (2.12) for MVJMs. Here, we consider each parameter which constructs $\boldsymbol{\Omega}$, in addition to the

baseline hazard, and illustrate how the approximation (3.1) is used to update parameter vector $\boldsymbol{\Omega}^{(m)} \to \boldsymbol{\Omega}^{(m+1)}$. For each parameter, we begin with the conditional expectation on the requisite log-likelihood (2.5)–(2.7) (i.e. the E-step), followed by the necessary steps taken to form its update (i.e. the M-step).

We continue with the shorthand notation for all expectations i.e. eschew their conditioning on the observed data and current parameter estimates with this held implicit going forward. In each presented update the parameter $\hat{x}$ does not represent the MLE per se, instead the estimate for the 'next iteration' $(m+1)$.

3.3.1 Update for D

The covariance matrix for the random effects D appears only in (2.5). We therefore have

$$\begin{aligned}\mathbb{E}_i[\log f(\boldsymbol{b}_i|\mathrm{D})] &= \mathbb{E}_i\left[-\frac{q}{2}\log 2\pi - \frac{1}{2}|\mathrm{D}| - \frac{1}{2}\boldsymbol{b}_i^\top \mathrm{D}^{-1}\boldsymbol{b}_i\right], \\ &= -\frac{q}{2}\log 2\pi + \frac{1}{2}|\mathrm{D}^{-1}| - \frac{1}{2}\mathrm{Tr}\Big\{\mathrm{D}^{-1}\mathbb{E}_i\left[\boldsymbol{b}_i\boldsymbol{b}_i^\top\right]\Big\},\end{aligned}$$

where $\mathrm{Tr}\{\cdot\}$ denotes the trace of its matrix argument. Since D is symmetric, the partial derivative with respect to its inverse is

$$\begin{aligned}\frac{\partial \mathbb{E}_i[\log f(\boldsymbol{b}_i|\mathrm{D})]}{\partial \mathrm{D}^{-1}} &= \frac{1}{2}\mathrm{D} - \frac{1}{2}\mathbb{E}_i\left[\boldsymbol{b}_i\boldsymbol{b}_i^\top\right], \\ \implies \hat{\mathrm{D}} &= \mathbb{E}_i\left[\boldsymbol{b}_i\boldsymbol{b}_i^\top\right].\end{aligned}$$

Now, we make use of the normal approximation (3.1), which tells us that $\mathbb{E}_i[\boldsymbol{b}_i] = \hat{\boldsymbol{b}}_i$ and $\mathrm{Var}[\boldsymbol{b}_i] = \hat{\Sigma}_i$. Trivially then we obtain

$$\begin{aligned}\mathrm{Var}[\boldsymbol{b}_i] = \hat{\Sigma}_i &= \mathbb{E}_i\left[\boldsymbol{b}_i\boldsymbol{b}_i^\top\right] - \mathbb{E}_i[\boldsymbol{b}_i]\,\mathbb{E}_i[\boldsymbol{b}_i]^\top, \\ \implies \mathbb{E}_i\left[\boldsymbol{b}_i\boldsymbol{b}_i^\top\right] &= \hat{\Sigma}_i + \mathbb{E}_i[\boldsymbol{b}_i]\,\mathbb{E}_i[\boldsymbol{b}_i]^\top, \\ &= \hat{\Sigma}_i + \hat{\boldsymbol{b}}_i\hat{\boldsymbol{b}}_i^\top.\end{aligned}$$

And finally we obtain the update

$$\hat{\mathrm{D}} = \frac{\sum_{i=1}^n \hat{\Sigma}_i + \hat{\boldsymbol{b}}_i\hat{\boldsymbol{b}}_i^\top}{n}, \tag{3.7}$$

which notably, owing to the approximation (3.1), does *not* require any integration.

3.3.2 Update for $\boldsymbol{\beta}$

The fixed effects $\boldsymbol{\beta}$ are housed in the log-likelihood for the longitudinal processes (2.7). Subject i's contribution to the log-likelihood is denoted by $\ell_i(\cdot)$. The log-likelihood with respect to $\boldsymbol{\beta}$ is

$$\ell_i(\boldsymbol{\beta}) \underset{\boldsymbol{\beta}}{\propto} -\frac{1}{2}(\boldsymbol{Y}_i - \boldsymbol{\eta}_i)^\top \mathrm{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\eta}_i)$$

where $\boldsymbol{\eta}_i = \mathrm{X}_i\boldsymbol{\beta} + \mathrm{Z}_i\boldsymbol{b}_i$. This has expected value

$$\begin{aligned}\mathbb{E}_i[\ell_i(\boldsymbol{\beta})] &= -\frac{1}{2}\mathbb{E}_i\Big[(\boldsymbol{Y}_i - \boldsymbol{\eta}_i)^\top \mathrm{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\eta}_i)\Big]\,, \\ &= -\frac{1}{2}\mathrm{Tr}\Big\{\mathrm{V}_i^{-1}\mathbb{E}_i\Big[(\boldsymbol{Y}_i - \boldsymbol{\eta}_i)(\boldsymbol{Y}_i - \boldsymbol{\eta}_i)^\top\Big]\Big\}\,,\end{aligned}$$

which we can appraise using the approximation (3.1). We set

$$\boldsymbol{\eta}_i = \mathrm{X}_i\boldsymbol{\beta} + \mathrm{Z}_i\boldsymbol{b}_i \overset{\text{appx.}}{\sim} N\Big(\mathrm{X}_i\boldsymbol{\beta} + \mathrm{Z}_i\hat{\boldsymbol{b}}_i, \mathrm{Z}_i\hat{\Sigma}_i\mathrm{Z}_i^\top\Big) = N(\hat{\boldsymbol{\mu}}_i, \mathrm{A}_i)\,, \tag{3.8}$$

where A_i is the variance-covariance matrix of the approximated (multivariate) normal distribution; its computation analogous to $\mathrm{Var}[aX + b] = a^2\mathrm{Var}[X]$ in the univariate case. Subsequently we define $\boldsymbol{\tau}_i = \mathrm{diag}(\mathrm{A}_i)^{1/2}$ and write the approximated conditional expectation

$$\tilde{\mathbb{E}}_i[\ell_i(\boldsymbol{\beta})] = -\frac{1}{2}\sum_{l=1}^{\varrho} w_l \mathrm{Tr}\Big\{\mathrm{V}_i^{-1}\Big[(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i - \boldsymbol{\tau}_i v_l)(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i - \boldsymbol{\tau}_i v_l)^\top\Big]\Big\}\,,$$

here $w_l,\ v_l,\ l = 1, \ldots, \varrho$ are the Gauss-Hermite weights and abscissae, respectively; $\sum_{l=1}^{\varrho} w_l = 1,\ \sum_{l=1}^{\varrho} v_l = 0$. In practise these weights and abscissae are found using `gauss.quad.prob` from `R` package `statmod` (Smyth, 2005).

The derivative of this approximate expectation with respect to $\boldsymbol{\beta}$ is

$$\frac{\partial\tilde{\mathbb{E}}_i[\ell_i(\boldsymbol{\beta})]}{\partial\boldsymbol{\beta}} = \mathrm{X}_i^\top\mathrm{V}_i^{-1}\sum_{l=1}^{\varrho} w_l(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i - \boldsymbol{\tau}_i v_l)\,,$$

where we note since $\sum_{l=1}^{\varrho} v_l = 0$ the derivative is evaluated at $\hat{\boldsymbol{b}}_i$ only. Equating the above to zero (with quadrature now eschewed) and equating to zero for all $i = 1, \ldots, n$ we obtain the closed-form update for $\boldsymbol{\beta}$

$$\mathrm{X}_i^\top\mathrm{V}_i^{-1}\Big(\boldsymbol{Y}_i - \mathrm{X}_i\boldsymbol{\beta} - \mathrm{Z}_i\hat{\boldsymbol{b}}_i\Big) = \boldsymbol{0} \implies \mathrm{X}_i^\top\mathrm{V}_i^{-1}\mathrm{X}_i\boldsymbol{\beta} = \mathrm{X}_i^\top\mathrm{V}_i^{-1}\Big(\boldsymbol{Y}_i - \mathrm{Z}_i\hat{\boldsymbol{b}}_i\Big)\,;$$

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \mathrm{X}_i^\top \mathrm{X}_i \right)^{-1} \left(\sum_{i=1}^{n} \mathrm{X}_i^\top \left[\boldsymbol{Y}_i - \mathrm{Z}_i \hat{\boldsymbol{b}}_i \right] \right). \tag{3.9}$$

3.3.3 Update for $\sigma^2_{\varepsilon_k}$

Once more we consider the longitudinal log-likelihood (2.7). We consider now its form for subject i *at time-point* j for ease of derivation i.e. considering the residual variance itself rather than its matrix representation V_{ik}. We additionally also approach this update for each $k = 1, \ldots, K$ independently, but note that the response identifier k is dropped in derivations for notational convenience. The log-likelihood contribution for subject i at timepoint j is

$$\ell_i\big(\sigma^2_\varepsilon\big) \underset{\sigma^2_\varepsilon}{\propto} -\frac{1}{2} \log \sigma^2_\varepsilon - \frac{1}{2\sigma^2_\varepsilon} (\mathrm{Y}_{ij} - \eta_{ij})^2 \,,$$

which has expected value

$$\mathbb{E}_i\big[\ell_i\big(\sigma^2_\varepsilon\big)\big] = -\frac{1}{2} \log \sigma^2_\varepsilon - \frac{1}{2\sigma^2_\varepsilon} \mathbb{E}_i\Big[(\mathrm{Y}_{ij} - \eta_{ij})^2\Big] \,.$$

Utilising the normal approximation (3.1) we set

$$\eta_{ij} = \mathrm{X}_{ij}\boldsymbol{\beta} + \mathrm{Z}_{ij}\boldsymbol{b}_i \overset{\text{appx.}}{\sim} N\Big(\mathrm{X}_{ij}\boldsymbol{\beta} + \mathrm{Z}_{ij}\hat{\boldsymbol{b}}_i, \mathrm{Z}_{ij}\hat{\Sigma}_i\mathrm{Z}_{ij}^\top\Big) = N\big(\hat{\mu}_{ij}, \tau^2_{ij}\big) \,,$$

and subsequently form the approximated expectation

$$\tilde{\mathbb{E}}_i\big[\ell_i\big(\sigma^2_\varepsilon\big)\big] = -\frac{1}{2} \log \sigma^2_\varepsilon - \frac{1}{2\sigma^2_\varepsilon} \sum_{l=1}^{\varrho} w_l (\mathrm{Y}_{ij} - \hat{\mu}_{ij} - \tau_{ij} v_l)^2 \,,$$

with derivative with respect to σ^2_ε

$$\frac{\partial \tilde{\mathbb{E}}_i\big[\ell_i\big(\sigma^2_\varepsilon\big)\big]}{\partial \sigma^2_\varepsilon} = -\frac{1}{2\sigma^2_\varepsilon} + \frac{1}{2\sigma^4_\varepsilon} \sum_{l=1}^{\varrho} w_l (\mathrm{Y}_{ij} - \hat{\mu}_{ij} - \tau_{ij} v_l)^2 \,.$$

Equating to zero and solving for σ^2_ε across all m_i observed time-points for all n subjects we obtain the closed-form update for σ^2_ε

$$\hat{\sigma}^2_\varepsilon = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{l=1}^{\varrho} w_l (\mathrm{Y}_{ij} - \hat{\mu}_{ij} - \tau_{ij} v_l)^2}{\sum_{i=1}^{n} m_i}. \tag{3.10}$$

3.3.4 Update for λ_0

We begin by rewriting the log-likelihood for the event-time process shown in (2.6) to eschew the integrand. This integration over the survival process is replaced with a finite summation over the process evaluated at the unique failure times, since the non-parametric estimator of baseline hazard is zero *except* at observed failure times (Henderson et al., 2000). Instead, we introduce matrices and vectors which contain all required information and functions of time to allow for said summation to occur easily. We now write

$$\begin{aligned}\log f(T_i, \Delta_i | \boldsymbol{b}_i; \boldsymbol{\Omega}) = \Delta_i \log \lambda_0(T_i) + \Delta_i \left[\boldsymbol{S}_i^\top \boldsymbol{\zeta} + \sum_{k=1}^{K} \gamma_k \boldsymbol{F}_{ik}^\top \boldsymbol{b}_{ik} \right] - \\ \lambda_0(\boldsymbol{u}_i)^\top \exp\left\{ \mathrm{S}_i \boldsymbol{\zeta} + \sum_{k=1}^{K} \gamma_k \mathrm{F}_{\boldsymbol{u}_i k} \boldsymbol{b}_{ik} \right\},\end{aligned} \tag{3.11}$$

where we introduce multiple items for notational convenience and brevity. The vector of failure times survived by subject i is denoted by $\boldsymbol{u}_i$. Vector $\boldsymbol{F}_{ik}$ denotes $\boldsymbol{W}_k(t)$ evaluated at T_i; if response k is to be modelled by an intercept-and-slope random effects specification then $\boldsymbol{F}_{ik} = (1, T_i)^\top$. The matrix $\mathrm{F}_{\boldsymbol{u}_i k}$ is defined in a similar spirit, except its *rows* are determined by $\boldsymbol{W}_k(t)$ evaluated at each element of $\boldsymbol{u}_i$. For convenience's sake we have also introduced S_i, simply representing a $\mathrm{len}(\boldsymbol{u}_i) \times p_s$ matrix, whose rows are replicates of $\boldsymbol{S}_i$.

With our rewritten log-likelihood established above, we can consider the conditional expectation on $\lambda_0(\cdot)$:

$$\begin{aligned}\ell_i(\lambda_0) &\underset{\lambda_0}{\propto} \Delta_i \log \lambda_0(T_i) - \lambda_0(\boldsymbol{u}_i)^\top \exp\left\{ \mathrm{S}_i \boldsymbol{\zeta} + \sum_{k=1}^{K} \gamma_k \mathrm{F}_{\boldsymbol{u}_i k} \boldsymbol{b}_{ik} \right\}, \\ \mathbb{E}_i[\ell_i(\lambda_0)] &= \Delta_i \log \lambda_0(T_i) - \lambda_0(\boldsymbol{u}_i)^\top \mathbb{E}_i \left[\exp\left\{ \mathrm{S}_i \boldsymbol{\zeta} + \sum_{k=1}^{K} \gamma_k \mathrm{F}_{\boldsymbol{u}_i k} \boldsymbol{b}_{ik} \right\} \right].\end{aligned}$$

To evaluate the expectation $\mathbb{E}_i\left[\exp\left\{\mathrm{S}_i\boldsymbol{\zeta} + \sum_{k=1}^{K} \gamma_k \mathrm{F}_{\boldsymbol{u}_i k} \boldsymbol{b}_{ik}\right\}\right]$, we make use of normal approximation (3.1)

$$\mathrm{S}_i \boldsymbol{\zeta} + \sum_{k=1}^{K} \gamma_k \mathrm{F}_{\boldsymbol{u}_i k} \boldsymbol{b}_{ik} \overset{\text{appx.}}{\sim} N\left(\mathrm{S}_i \boldsymbol{\zeta} + \sum_{k=1}^{K} \gamma_k \mathrm{F}_{\boldsymbol{u}_i k} \hat{\boldsymbol{b}}_{ik}, \mathrm{Q} \hat{\Sigma}_i \mathrm{Q}^\top \right) = N(\hat{\boldsymbol{\mu}}_i, \mathrm{A}_i), \tag{3.12}$$

with A_i previously defined and the matrix $\mathrm{Q} = \mathrm{F}_{\boldsymbol{u}_i} \mathrm{diag}(\boldsymbol{\gamma}^*)$, where $\boldsymbol{\gamma}^*$ is the vector with q_k replicates of $\gamma_k \; \forall \; k = 1, \ldots, K$ and $\mathrm{F}_{\boldsymbol{u}_i}$ is the horizontal concatenation of $\mathrm{F}_{\boldsymbol{u}_i 1}, \ldots, \mathrm{F}_{\boldsymbol{u}_i K}$.

We can then approximate the expectation above as

$$\tilde{\mathbb{E}}_i[\ell_i(\lambda_0)] = \Delta_i \log \lambda_0(T_i) - \lambda_0(\boldsymbol{u}_i)^\top \sum_{l=1}^{\varrho} w_l \exp\{\hat{\boldsymbol{\mu}}_i + \boldsymbol{\tau}_i v_l\}\,.$$

Finally, we can take derivative with respect to the baseline hazard, equate to zero, and fairly trivially form the update for $\hat{\lambda}_0(\cdot)$:

$$\hat{\lambda}_0(u) = \frac{\sum_{i=1}^{n} \Delta_i I(T_i = u)}{\sum_{i=1}^{n} \sum_{l=1}^{\varrho} w_l \exp\{\hat{\boldsymbol{\mu}}_i + \boldsymbol{\tau}_i v_l\}\, I(T_i \geq u)}. \tag{3.13}$$

3.3.5 Update for survival parameters $\boldsymbol{\Phi}$

Lastly, we consider the update for the pair of survival parameters $\boldsymbol{\Phi} = \left(\boldsymbol{\gamma}^\top, \boldsymbol{\zeta}^\top\right)^\top$. Utilising the same rewritten survival log-likelihood (3.11) we can form the conditional expectation

$$\begin{aligned}
\ell_i(\boldsymbol{\Phi}) &\underset{\boldsymbol{\Phi}}{\propto} \Delta_i \left\{ \boldsymbol{S}_i^\top \boldsymbol{\zeta} + \sum_{k=1}^{K} \gamma_k \boldsymbol{F}_{ik}^\top \boldsymbol{b}_{ik} \right\} - \lambda_0(\boldsymbol{u}_i)^\top \exp\left\{ \mathrm{S}_i \boldsymbol{\zeta} + \sum_{k=1}^{K} \gamma_k \mathrm{F}_{\boldsymbol{u}_i k} \boldsymbol{b}_{ik} \right\}, \\
\mathbb{E}_i[\ell_i(\boldsymbol{\Phi})] &= \Delta_i \left\{ \boldsymbol{S}_i^\top \boldsymbol{\zeta} + \sum_{k=1}^{K} \gamma_k \boldsymbol{F}_{ik}^\top \mathbb{E}_i[\boldsymbol{b}_{ik}] \right\} - \lambda_0(\boldsymbol{u}_i)^\top \mathbb{E}_i\left[\exp\left\{ \mathrm{S}_i \boldsymbol{\zeta} + \sum_{k=1}^{K} \gamma_k \mathrm{F}_{\boldsymbol{u}_i k} \boldsymbol{b}_{ik} \right\} \right]
\end{aligned}$$

where from (3.1) we know $\mathbb{E}_i[\boldsymbol{b}_{ik}] = \hat{\boldsymbol{b}}_{ik}$ and $\mathbb{E}_i\left[\exp\left\{ \mathrm{S}_i \boldsymbol{\zeta} + \sum_{k=1}^{K} \gamma_k \mathrm{F}_{\boldsymbol{u}_i k} \boldsymbol{b}_{ik} \right\}\right]$ is approximated in the same way as (3.12). We can then form the approximated expectation

$$\tilde{\mathbb{E}}_i[\ell_i(\boldsymbol{\Phi})] = \Delta_i \left\{ \boldsymbol{S}_i^\top \boldsymbol{\zeta} + \sum_{k=1}^{K} \gamma_k \boldsymbol{F}_{ik}^\top \hat{\boldsymbol{b}}_{ik} \right\} - \lambda_0(\boldsymbol{u}_i)^\top \sum_{l=1}^{\varrho} w_l \exp\{\hat{\boldsymbol{\mu}}_i + \boldsymbol{\tau}_i v_l\}\,. \tag{3.14}$$

Because the elements of $\boldsymbol{\Phi}$ are housed within an exponent, a closed-form update doesn't exist. Instead, we consider the one-step Newton-Raphson update as presented in (2.12). Here we note that because $\boldsymbol{\gamma}$ is effectively found 'in' the $\boldsymbol{\tau}_i$ term, differentiation of the approximated expectation above is an unattractive task. With this in mind, we use numerical differentiation techniques to calculate both the (profile) score $s_i(\boldsymbol{\Phi})$ and Hessian $\mathrm{H}_i(\boldsymbol{\Phi})$ of the conditional expectation above. Additionally, we note that in the above we substitute $\lambda_0(\cdot)$ with $\hat{\lambda}_0(\cdot)$ given by (3.13).

3.4 Simulation studies

With the methodology established in Sections 3.2 and 3.3, we endeavour to ascertain the parameter estimation capabilities of the approximate EM algorithm for MVJMs. Moreover, we consider several simulation scenarios with overarching goal of establishing said

capabilities, along with how the method is affected by different data structures and parameter values. Due to the large quantity of simulations we consider, we first define the 'standard' simulation scenario. Unless otherwise stated, we consider $N = 100$ simulations for each scenario.

3.4.1 The 'standard' simulation scenario

We seek to define the *default* set of simulations parameters we use. This allows us to enjoy perquisite brief explanations of all future simulations we consider by e.g. pointing out what's changed from said default set.

We set the number of subjects $n = 500$, each of whom have their k^{th} continuous (and Gaussian) longitudinal measurements taken at the regularly-spaced vector of possible times $\boldsymbol{t} = (0, \ldots, \kappa)^\top$ where $\kappa = 5$ denotes the maximal follow-up time. The vector $\boldsymbol{t}$ has length r, such that at as r increases these follow-up measurement times become more regular i.e. $\boldsymbol{t}$ is more 'dense'. The fixed effects are defined by $\boldsymbol{\beta}_{\text{odd}} = (2, -0.1, 0.1, -0.2)^\top$ and $\boldsymbol{\beta}_{\text{even}} = -\boldsymbol{\beta}_{\text{odd}}$; the elements correspond to an intercept term, time $\boldsymbol{t}$ and two baseline covariates in the form of standard normal realization x_{i1} and a single Bernoulli draw ($p = 0.5$), x_{i2}. This specification of X_{ik} matches Section 2.5.1 and is the same for $k = 1, \ldots, K$. We simulate under a random intercept and slope specification unless otherwise stated, subsequently defining the $q \times q = 2K \times 2K$ variance-covariance matrix for the random effects $\mathrm{D} = \bigoplus_{k=1}^{K} \mathrm{diag}(0.25, 0.09)$, with correlation induced between random intercept terms by setting $\mathrm{D}_{ef} = 0.125 \; e, f = 1, 3, \ldots, q-1, \; e \neq f$. We set residual variance $\sigma^2_{\varepsilon_k} = 0.16 \; \forall \; k = 1, \ldots, K$.

Survival times T_i, along with an independent potential censor-time C_i, are simulated for each subject using the methodology outlined in Section 2.5.3 and subsequent failure indicators Δ_i are generated. We set only a single time-invariant coefficient ζ which is attached to x_{i2} in the survival sub-model. The association parameters are set as $\gamma_{\text{odd}} = 0.5$ and $\gamma_{\text{even}} = -0.5$. The shape, α, and (log) scale, $\log \nu$, of the Gompertz baseline hazard we introduced in Section 2.5.2 are set to $\alpha = 0.1, \; \log \nu = -0.3$ by default, which results in an average failure rate of approximately 30% over *all* N simulated datasets. The resultant distribution of failure times then resembling the corresponding flatter distributions seen in Section 2.6.1. The censoring rate is held at $\Upsilon = e^{-3.5}$ which results in approximately 13% censoring. We then follow the truncation procedure outlined in Section 2.5.3 for all subjects to obtain $\boldsymbol{t}_i \; \forall \; i = 1 \ldots, n$.

Our standard joint model simulation set-up for the k^{th} longitudinal response for subject i

at time j is

$$\begin{cases} y_{ikj} &= (\beta_{k0} + b_{ik0}) + (\beta_{k1} + b_{ik1})\, t_{ij} + \beta_{k2} x_{i1} + \beta_{k3} x_{i2} + \varepsilon_{ikj} \\ \lambda_i(t) &= \lambda_0(t) \exp\Big\{ x_{i2}\zeta + \sum_{k=1}^{K} \gamma_k (b_{ik0} + b_{ik1} t) \Big\}, \end{cases} \tag{3.15}$$

with

$$\boldsymbol{b}_i \sim N_q(\mathbf{0}, \mathrm{D}), \qquad \varepsilon_{ikj} \sim N\big(0, \sigma^2_{\varepsilon_k}\big), \qquad \boldsymbol{b}_{ik} \perp\!\!\!\perp \boldsymbol{\varepsilon}_{ik}.$$

With the default simulation scenario established, we proceed with several simulation studies which aim to monitor multiple facets of estimation via the approximate EM algorithm.

All simulation studies and applications in the thesis were executed on an Ubuntu Desktop with 4GHz Intel core i7 with 8GB RAM using `R` version 4.2.0. No high performance computing facility was used. All `R` code is available online: `https://github.com/jamesmurray7/thesis`.

3.4.2 Sample size n

We first consider the impact of varying sample size n. We select four (fairly arbitrary) choices, setting $n \in \{100, 250, 500, 1000\}$. Joint models are fit to the largest simulated sample sizes $n = 1000$ using stopping criterion $\xi_2 = 0.01$ in (3.4). We present estimates for the survival parameters, $\hat{\boldsymbol{\Phi}} = \left(\hat{\boldsymbol{\gamma}}^\top, \hat{\zeta}\right)^\top$, from all $N = 100$ simulations at each of the four sample size choices in Figure 3.1. Immediately, we note that the true value (dashed line) is well estimated, with decreasing spread as the sample size increases.

In-depth results from this simulation study is presented in Appendix B.1.1. From these, we note that the algorithm generally produces very good point estimates for the true parameter value used in simulation, with empirical variability which decreases as the sample size increases. We note that the coverage of parameters appears to be conservative, most egregiously so for the smallest sample size; this levels off slightly at $n = 250$, and is more palatable for the slightly larger sample sizes $n \in \{500, 1000\}$. This phenomena is likely related in part to the discussions surrounding power given in Appendix A.3. Especially for $n = 100$ where the confidence intervals for estimates are very wide given the lack of information in the data and the estimates themselves more likely to be influenced by random variability amongst the simulated data; as n increases the ability to detect the true parameters $\boldsymbol{\Omega}$ also improves, with minimal bias observed amongst parameters for $n = 1000$.

Finally, the time taken for the approximate EM algorithm to converge and calculation of standard errors unsurprisingly increases with n. For the smallest sample size, this elapsed time was (median [IQR]) 1.799 [1.452, 2.479] seconds, approximately doubling for

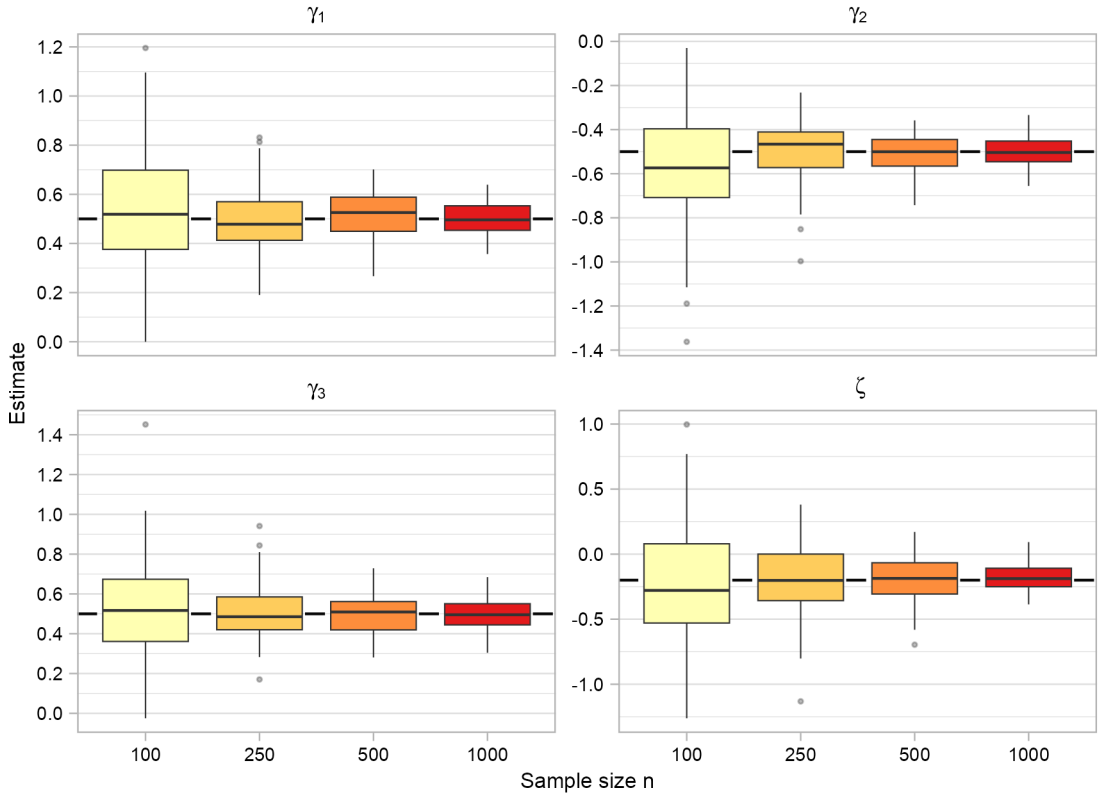


Figure 3.1: Estimates for survival parameters $\hat{\boldsymbol{\gamma}}$ and $\hat{\zeta}$ for different simulated sample sizes n. The dashed line signifies the true parameter value.

$n = 250$ with 3.553 [3.314, 3.769] seconds; at $n = 500$ 8.525 [8.243, 8.886] seconds and a larger 'jump' at the largest sample size 26.962 [26.193, 27.591] seconds. We note that one should not expect *proportional* increases in computation time, with this large increase in time taken largely attributable to the sheer number of `optim` steps (i.e. step 2. in Section 3.2.4) as well as the size of survival design matrices (3.11) which is determined by the number of unique failure times. The median number of EM iterations per second completed by the algorithm decreases from 5.0 for $n = 100$, 1.8 for $n = 250$, 0.73 for $n = 500$ and 0.20 for $n = 1000$. Interestingly however, the $n = 1000$ case systematically requires fewer iterations to achieve convergence; potentially due to the greater amount of data providing more information for the algorithm to provide better estimates for the conditional expectations housed in the E-step.

3.4.3 Number of longitudinal responses K

The main 'selling point' of the approximate EM algorithm is that irregardless of the dimension of the random effects, the conditional expectations are appraised against a

univariate normal distribution. Therefore, we would expect to observe a non-exponential increase in computation times as we increase the number of longitudinal responses (i.e. increasing the dimension of random effects), which one would observe under traditional existing methodologies (which we explore in Section 3.5).

We elect $K \in \{1, 2, 3, 5, 7\}$, and for each simulation scenario we set the variance-covariance matrix to be the diagonal matrix $\mathrm{D} = \bigoplus_{k=1}^{K} \mathrm{diag}(0.25, 0.06)$. A swathe of parameter estimates are presented in Appendix B.1.2. Here we note, particularly for larger K, that the parameter estimates are slightly conservative in nature, this is likely due to inadequate sample size n for the complexity of the model determined by the number of parameters (i.e. treating $\gamma_k,\ k = 1, \ldots, K$ as time-varying covariates), with the perhaps relatively small $n = 500$ in combination with the relatively low event rate, resulting in wider confidence intervals than one may expect in an appropriately-powered empirical study; see Appendix A.3. The focal point of the simulations we carried out was largely from a computation time-standpoint.

The computation times are perhaps best captured in Figure 3.2, where we note an (approximately) linear increase in computation times as the dimension of random effects increases from $q = 2 \to 4 \to \cdots \to 14$.

3.4.4 Length of follow-up period r

Given the underlying normal approximation utilised in Sections 3.2–3.3, we expect presence of more data in the longitudinal sub-model to improve parameter estimate performance. With this in mind, we select four different candidate values for the *maximal* follow-up period $r \in \{3, 5, 10, 15\}$. We note in these simulations that the average failure time is similar. The parameter estimates for the survival parameters $\hat{\mathbf{\Phi}}$ from all simulations at each r is shown in Figure 3.3, with detailed tabulation provided in Appendix B.1.3.

Results from Rizopoulos (2012a) as well as Baghishani and Mohammadzadeh (2012) indicate that as the minimal observed profile length increases, the bias in parameter estimates should decrease (and performance in general improve), Figure C.1 shows the bias in each of the parameters estimated by the joint model. Generally speaking, performance is noticeably worse at $r = 3$, which quickly improves even for $r = 5$, with there being no real distinguishing features as r increases, especially between $r = 10$ and 15, save for the average estimated standard error approaching the true empirical standard deviation of parameter estimates, indicating that the approximation appears to better capture the true variability of $\hat{\mathbf{\Omega}}$ as a longer longitudinal profile becomes available.

Remark. We note (particularly in Table B.7 in Appendix B.1.3) that the improvement

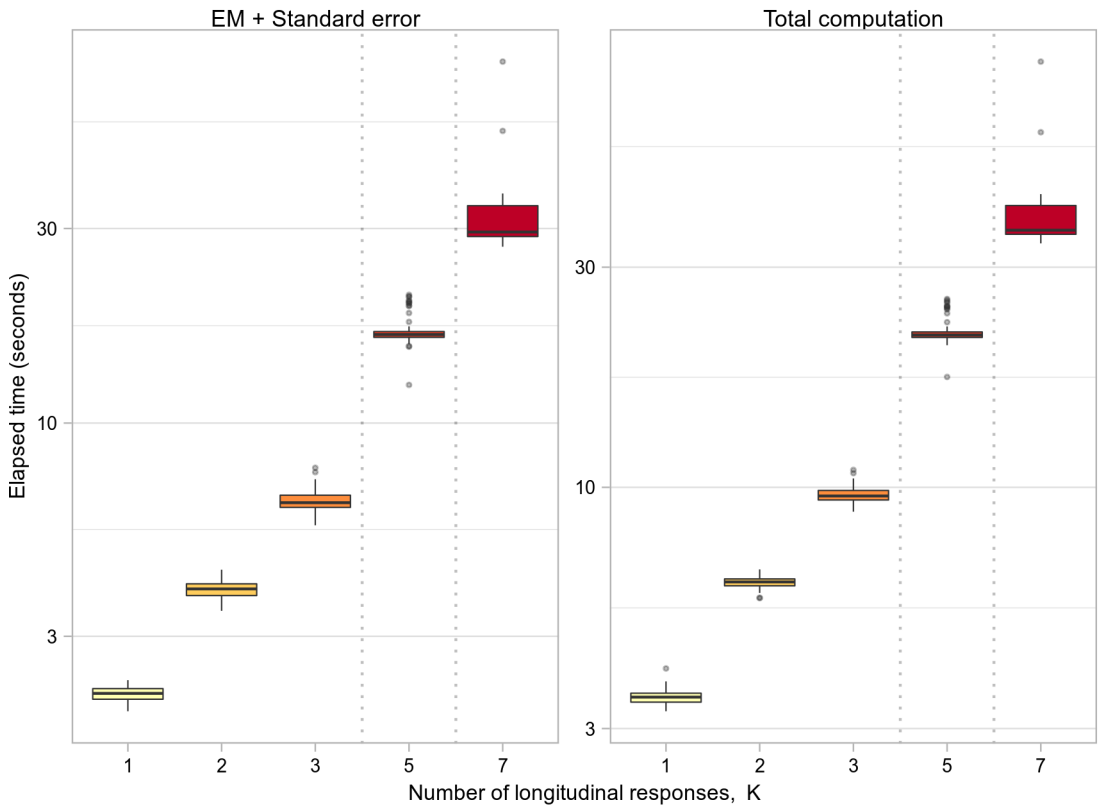


Figure 3.2: Elapsed time (seconds) for convergence of the approximate EM algorithm (+ calculation of standard errors), as well as total computation time for choices of K. Vertical dotted lines represent breaks in the x-axis. A $\log_{10}$ scale is used between ticks on the y-axis.

isn't perhaps as marked as one may expect; this could be due to $N = 100$ being a relatively low number of simulations to carry out (i.e. larger Monte Carlo error), and that the minimal (and median) value for m_i does not change dramatically between sets of r, which may not inherently lead to better-estimated random effects.

3.4.5 Failure rate ω

We now alter the two parameters which control the baseline hazard: The shape α and scale $\log \nu$, as outlined in Section 3.4.1. We hold the shape constant at its default value $\alpha = 0.1$, representing a gradually increasing underlying hazard, instead altering the scale parameter to achieve some arbitrarily-set average failure rate across the $N = 100$ generated sets of data. The values for $\log \nu$ and resulting (approximate) failure rates ω are:

$$\overset{\log \nu \,=}{\{-4.30, -3.00, -2.15, -1.40\}} \rightarrow \overset{\omega \,=}{\{10\%, 30\%, 50\%, 70\%\}}.$$

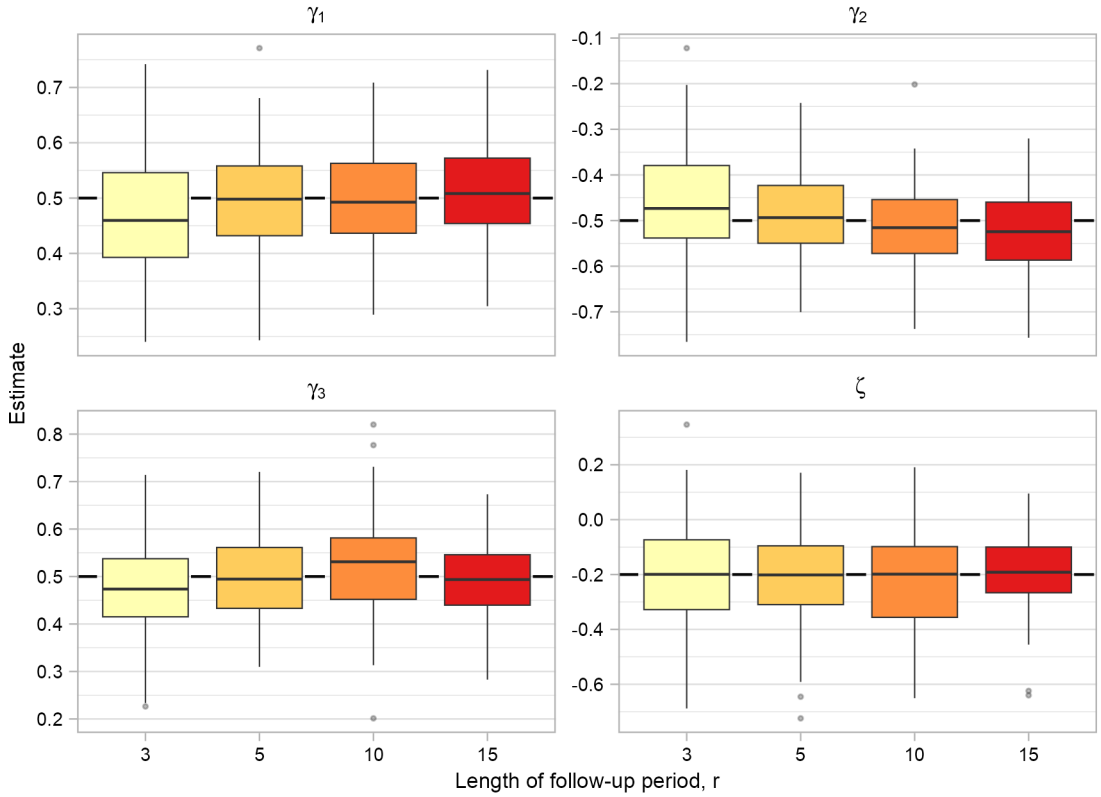


Figure 3.3: Estimates for survival parameters $\hat{\boldsymbol{\gamma}}$ and $\hat{\zeta}$ for different maximal longitudinal profile lengths r. The dashed line signifies the true parameter value.

We expect *a priori* that an increased number of failures will produce more 'concentrated' estimates for $\hat{\zeta}$ (simply due to there being more 'information'), with interest additionally falling on the performance of $\hat{\boldsymbol{\gamma}}$, with the other parameters $\hat{\boldsymbol{\Omega}}_{-\hat{\boldsymbol{\Phi}}}$ also of (secondary) interest. Estimates for survival parameters are presented in Figure 3.4, and detailed tabulation available in Appendix B.1.4 for all parameters.

Appraising first the N estimates for $\hat{\boldsymbol{\Phi}}$, we note generally the spread around the mean parameter estimate tends to decrease for all parameters, most obviously for the time-invariant $\hat{\zeta}$ with the lowest failure rate $\omega = 10\%$ producing the worst survival parameter estimates. We may then conclude from these results that the algorithm appears to perform best when the failure rate is in the region of 30–50%.

For the elements of $\text{vech}\big(\hat{\mathrm{D}}\big)$, we note little to distinguish between failure rates. Investigating the fixed effect parameters $\hat{\boldsymbol{\beta}}$, we note that the term attached to time $\hat{\beta}_{k1}$ enjoys lowest bias at $\omega = 10\%$, obviously since fewer truncations of the longitudinal profile occur with little distinguishing the remaining fixed effects in $\hat{\boldsymbol{\beta}}$. Finally, the residual variances $\hat{\sigma}^2_{\varepsilon_k}$ are well estimated irregardless of ω.

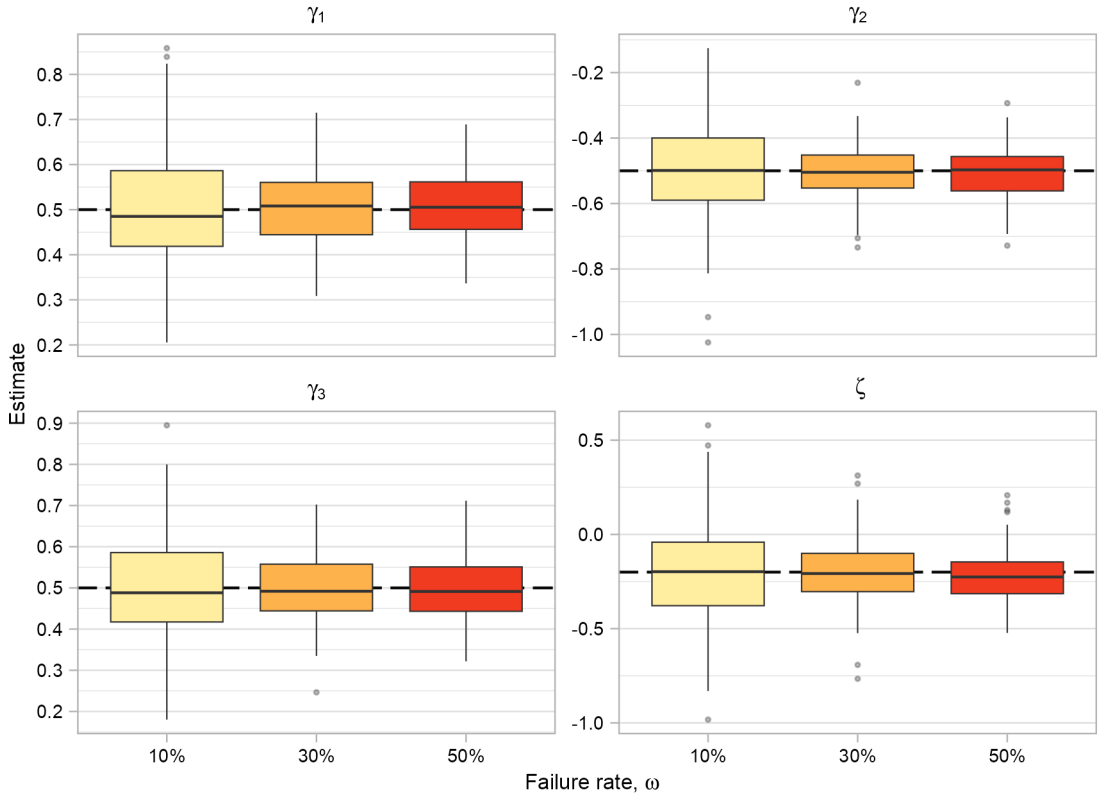


Figure 3.4: Estimates for survival parameters $\hat{\boldsymbol{\gamma}}$ and $\hat{\zeta}$ for different (approximate) failure rates ω. The dashed line signifies the true parameter value.

Remark. We note that the computation time increases as the proportion of the sample who fails increases. This is partially due to the simulation generating $\sum_{i=1}^{n} \Delta_i$ unique failure times, so that as ω increases, the matrices in (3.11) become much larger, thereby having a 'knock-on' effect on subsequent matrix algebra necessary in computations required in the E-step.

3.4.6 Magnitude of random effects variance

We now consider increasing the variances associated with random effects to be arbitrarily large. We alter *only* the diagonal elements, meaning that the off-diagonal 'default' values for matrix D we outline in Section 3.4.1 remain unchanged, inducing light-to-moderate correlation across random intercepts, which are not reported. We consider multiplicative increases on these diagonal elements. Namely, we introduce $\mathrm{D}^{(1)}$ and $\mathrm{D}^{(2)}$ which have $(e, f)^{\text{th}}$ element set as $\mathrm{D}^{(1)}_{e,f} = 3\mathrm{D}_{e,f}$, $\mathrm{D}^{(2)}_{e,f} = 10\mathrm{D}_{e,f}\ \forall\ e = f$ and $\mathrm{D}^{(1)}_{e,f} = \mathrm{D}^{(2)}_{e,f} = \mathrm{D}_{e,f}\ \forall\ e \neq f$ representing two candidate variance-covariance matrices with ratios of the random intercept to random slope variance held the same as the default D. We additionally introduce

the matrices $\mathrm{D}^{(3)}$ and $\mathrm{D}^{(4)}$ which once more share the same off-diagonal elements with D but with element-wise scale factors applied to their main diagonals:

$$\begin{aligned}\operatorname{diag}\left(\mathrm{D}^{(3)}\right) &= \operatorname{diag}(\mathrm{D}) \odot (3, 1.5, 5, 0.9, 4, 1.2)^\top, \quad \text{and} \\ \operatorname{diag}\left(\mathrm{D}^{(4)}\right) &= \operatorname{diag}(\mathrm{D}) \odot (10, 2, 13, 3, 15, 4)^\top,\end{aligned}$$

where $\odot$ denotes element-wise multiplication.

The estimates for survival parameters are presented in Figure 3.5 as well as in the detailed capabilities for all parameters in Appendix B.1.5. We note all parameters $\hat{\boldsymbol{\Omega}}_{-\hat{\boldsymbol{\Phi}}}$ appear to be well estimated no matter the characterisation of the variance-covariance matrix. For the association parameters $\hat{\boldsymbol{\gamma}}$ we note that performance is slightly worse under $\mathrm{D}^{(1)}$, wherein the random slope variance is large relative to that of the random intercept, which perhaps leads to muddied capabilities in estimating the time-varying parameter $\boldsymbol{\gamma}$. Likewise, under $\mathrm{D}^{(4)}$ – wherein the random slope variance is *small* compared to the random intercept's – we note similar phenomena, although in neither scenario is the performance poor. The association parameters $\hat{\boldsymbol{\gamma}}$ are best-estimated in this simulation study under $\mathrm{D}^{(2)}$ and $\mathrm{D}^{(3)}$ wherein one posits that the 'ratio' of random intercept-to-slope variance is more sensibly defined; allowing both the simulated data and the model to adequately establish simulated trajectories. We note the median computation time does not drastically differ amongst candidate D.

3.4.7 Closing remarks

Through the simulation studies carried out in Sections 3.4.2 – 3.4.6, we tweaked many of the simulation 'tuning knobs' we identified in Section 2.5 in an effort to create distinct simulation scenarios; visualisations of the central tendency of estimates for the survival parameters $\hat{\boldsymbol{\Phi}} = \left(\hat{\boldsymbol{\gamma}}^\top, \hat{\zeta}\right)^\top$ were presented throughout, with detailed tabulation of *all* parameter estimates $\hat{\boldsymbol{\Omega}}$ given in Appendix B.1. For all cases we noted that parameter estimation was not negatively affected by the changes made to the simulated data.

There is undoubtedly something of a trade-off between the parameter estimates – in certain simulation scenarios being deemed conservative – and the fast computation time. For the scenarios we considered in these simulation studies, we argue that this exchange is reasonable. Indeed, Bernhardt et al. (2015) posited that if more accurate parameter estimates, or standard errors, were required then the approximate EM algorithm as used here could be used instead to obtain starting values for e.g. an MCMC scheme in a relatively quick manner i.e. eschewing much of the burn-in phase, which may be arbitrarily long for complex joint models. We also refer to potential consternation surrounding sample size and incidence rate as outlined in Appendix A.3; the algorithm and/or the route taken

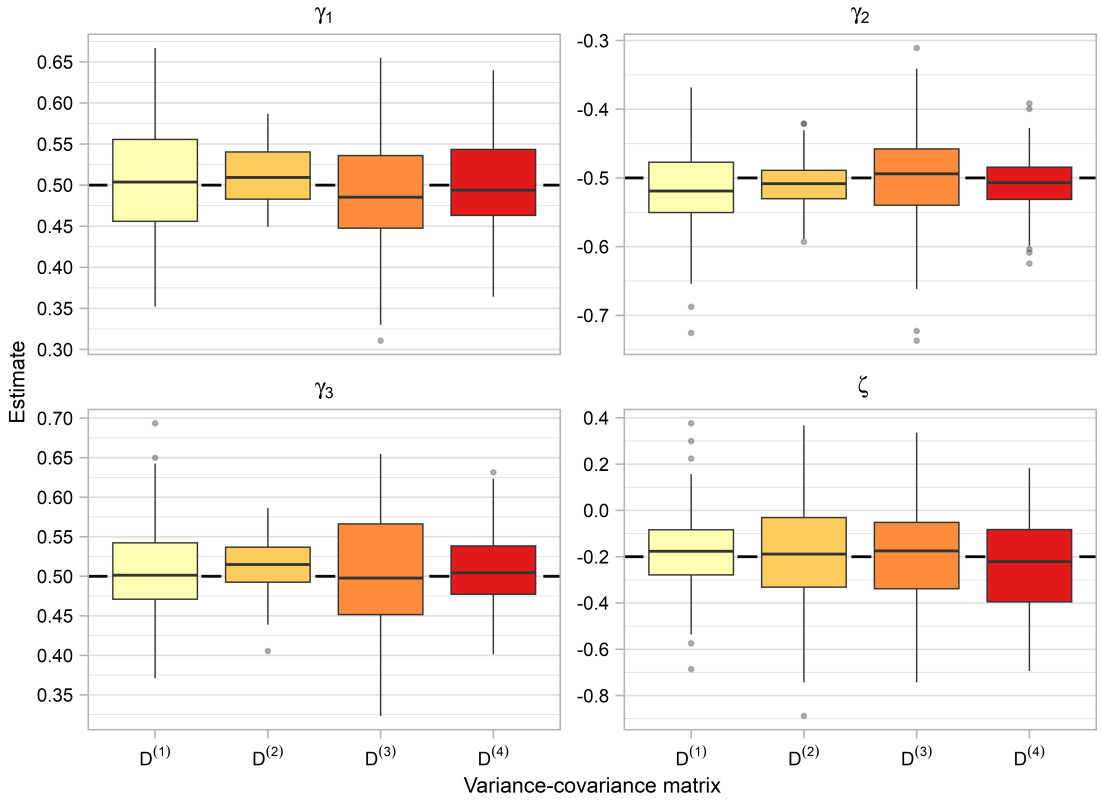


Figure 3.5: Estimates for survival parameters $\hat{\boldsymbol{\gamma}}$ and $\hat{\zeta}$ for different variance-covariance matrices D. The dashed line signifies the true parameter value. For the matrices $\mathrm{D}^{(1)}, \ldots, \mathrm{D}^{(4)}$ see Section 3.4.6.

to establish uncertainty in the parameter estimates as given in Sections 2.4.2 and 3.2.4, respectively, may suffer from these extraneous factors.

3.5 Comparison with existing software

We now seek to compare the parameter estimation capabilities of the approximate EM algorithm established in this chapter with existing methodology. We choose the software package `joineRML` (Hickey et al., 2018a), which fits multivariate joint models as explored in Chapter 2 by maximum likelihood using a Monte Carlo E-step, which we explored in Section 2.3.5. Here, we simply endeavour to show that the computation time spent in the EM algorithm follow different trajectories as the dimension of random effects increases.

We elect $K \in \{3, 5, 7\}$, and for each simulation scenario here we set the sample size $n = 250$ and the variance-covariance matrix to be the diagonal matrix $\mathrm{D} = \bigoplus_{k=1}^{K} \mathrm{diag}(0.25, 0.06)$. The main reason for setting the sample size at $n = 250$ is to avoid computation issues

under `joineRML` for largest K. A byproduct of this safeguard is that at largest K, we observe extremely conservative coverage for *both* methodologies, as both the model-fitting approaches produce large standard errors in circumstances where the sample size is small. As we discuss in Appendix A.3 (and as alluded to in Sections 3.4.2), one posits that a certain sample size and/or failure rate is required for a given number of longitudinal responses; with little research undertaken here in the multivariate joint modelling context. With this in mind, we do not present tabulated results, instead opting for graphical representation.

The routine undertaken by `joineRML` is slightly different to the approximate EM algorithm we use. Though both methods obtain initial conditions for $\boldsymbol{\Omega}$ via the process outlined in Section 3.2.2 – bar minor difference in external packages used – `joineRML` notably performs a *separate* EM algorithm on the K longitudinal processes only in order to start its MCEM procedure as close to the maximiser as possible; in particular the off-block-diagonal elements in D are 'better' than those obtained by the procedure in Section 3.2.2. With this in mind, we allow `joineRML` to undertake this 'pre-EM' EM algorithm until convergence, which we set a high tolerance of 0.1 on, so that the starting values are likely more in-line with those from the approximate EM algorithm: The aim here is to time the EM algorithm itself, and so this endeavour for similar starting values is undertaken; whereafter the time discrepancies observed are more likely to stem from computational differences mostly arising from the different numerical integration routines.

We use the same convergence criteria (3.4) across both methodologies, with tolerance values set as described in Section 3.2.3. Philipson et al. (2020) showed that usage of quasi Monte Carlo methods for sampling from $\boldsymbol{b}_i|\boldsymbol{Y}_i$ (2.18), namely the Sobol sequence, produced fastest fitting joint models out of `joineRML`'s available options, so is used in these comparative simulations. The default values for the number of MC iterations, scale factor increase in sample size for consecutive non-convergent iterations, and so on, are left at their default values i.e. we assume said values are preset at least owing in part to their performance.

We begin with perhaps the most important feature: Computation time. Figure 3.6 shows that, whilst neither approach appears to suffer from truly exponential increase in computation time, being able to shirk necessity of integration by Monte Carlo sampling allows the approximate EM algorithm to be faster as a baseline.

Next, we seek to compare the parameter estimation capabilities of the survival parameters $\hat{\boldsymbol{\Phi}}$ from each approach, which are presented in Figure 3.7. Here, perhaps the most striking conclusion is the similarity between the estimates. Though this is perhaps not surprising: The modal estimate for $\hat{\boldsymbol{b}}_i$ obtained in (3.2) is likely extremely close to the mean value for of the sampled collection of (some function of) $\boldsymbol{b}_i$ obtained by MCEM (2.28), i.e. the

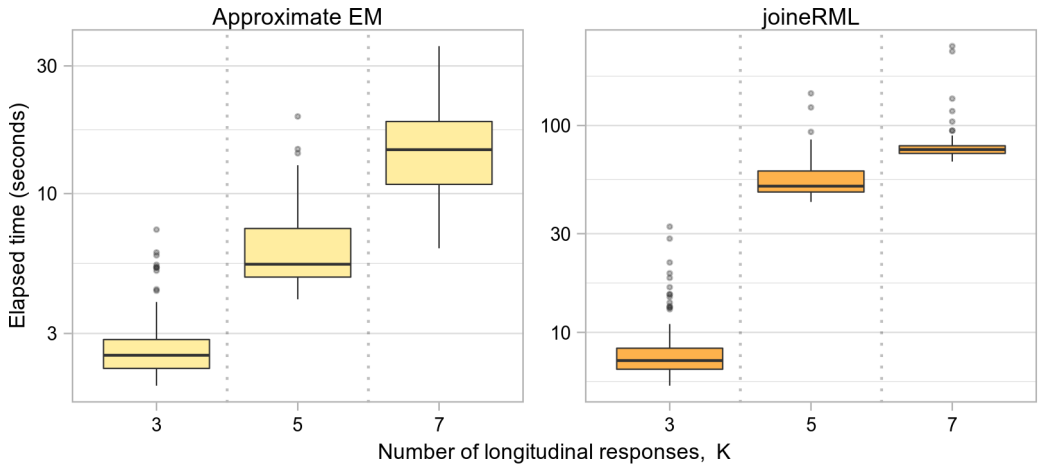


Figure 3.6: Elapsed time (seconds) for convergence of the approximate EM algorithm (+ calculation of standard errors) and convergence of the MCEM scheme in `joineRML`. Vertical dotted lines represent breaks in the x-axis. A $\log_{10}$ scale is used between ticks on the y-axis.

expectations required at each EM iteration are evaluated to be approximately the same across methods, some further investigation along this line of thought is given later in Section 4.7.

The parameter estimates themselves indicate that performance is good across both methods, with there being some evidence of overestimation in absolute value of some $\hat{\boldsymbol{\gamma}}$ terms (e.g. $\hat{\gamma}_4$ and $\hat{\gamma}_7$ in $K = 7$; $\hat{\gamma}_1$, $\hat{\gamma}_3$ and $\hat{\gamma}_5$ in $K = 5$). An interesting feature to note is that the spread of estimates for $\hat{\zeta}$ is consistent irregardless of the number of longitudinal responses; indicating that the variance around this time-invariant parameter is reliant on the (potential interaction of) sample size and failure rate.

The parameter estimates in Figure 3.7 are supplemented by graphical representation ('X') of the average Wald confidence intervals of these parameter estimates, as dictated by the estimated standard errors, in addition to the empirical variance captured by the boxplots already. These neatly capture the performance issue surrounding the method of standard error calculation (2.29), and as noted earlier a much larger sample size is likely needed to ameliorate this. However, as noted at the outset of this exercise, the simulations here are largely perfunctory: Establishing a handle on capabilities across methods with predominant focus on the computation time.

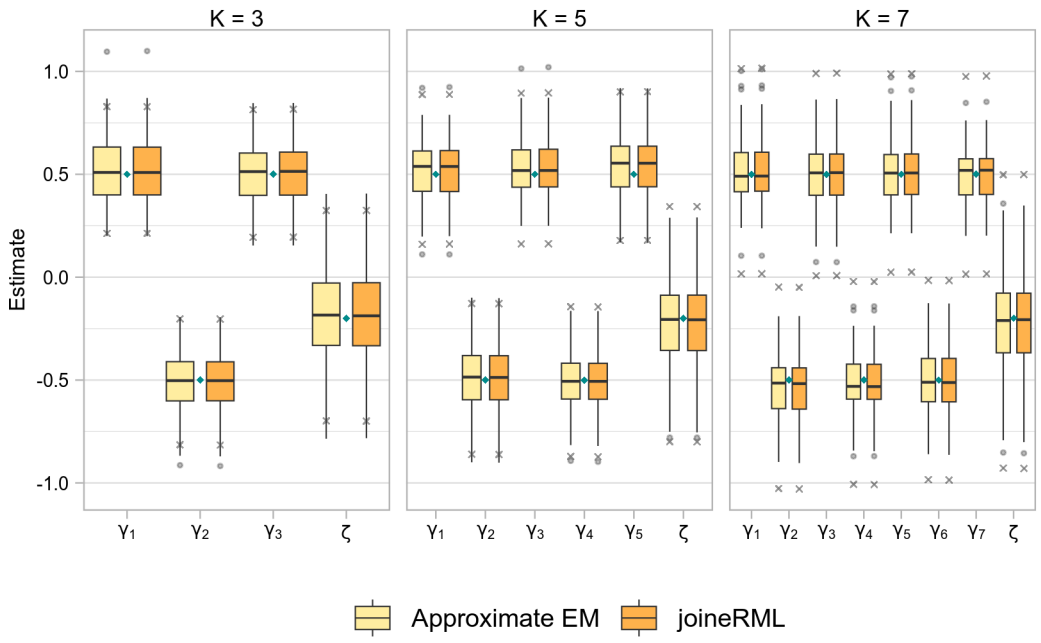


Figure 3.7: Estimates for survival parameters $\hat{\boldsymbol{\gamma}}$ and $\hat{\zeta}$ across 100 simulated joint model fits performed by the approximate EM algorithm and `joineRML`. A blue diamond denotes true value. In addition to the boxplots of point-estimates, grey crosses represent the average 2.5% and 97.5% point estimates across all corresponding simulations (i.e. the empirical estimated range based on the standard error).

3.6 Sensitivity analysis

3.6.1 t-distributed random effects

The random effects used in simulation have been realizations from a multivariate normal distribution with zero-mean and pre-specified variance-covariance matrix D. One could argue such an assumption is somewhat prohibitive: In reality there would likely be a number of deviations one would classify as outliers. A t_ν-distribution however would allow for larger amounts of variation in these random effects, particularly for lower degrees of freedom ν.

This avenue is perhaps doubly interesting since we consider $f(\boldsymbol{b}_i|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega})$ to be approximately normal and changing the true distribution of the random effects to themselves be non-normal could impact performance. We can therefore consider a brief simulation study to appraise how stringent this approximation is. A priori, we expect estimation capabilities of the variance-covariance matrix D used in the multivariate t_ν distribution to be poor for lower degrees of freedom ν. This is because internally, we estimate vech(D) using properties of the multivariate normal distribution, as outlined in Section 3.3.1. We

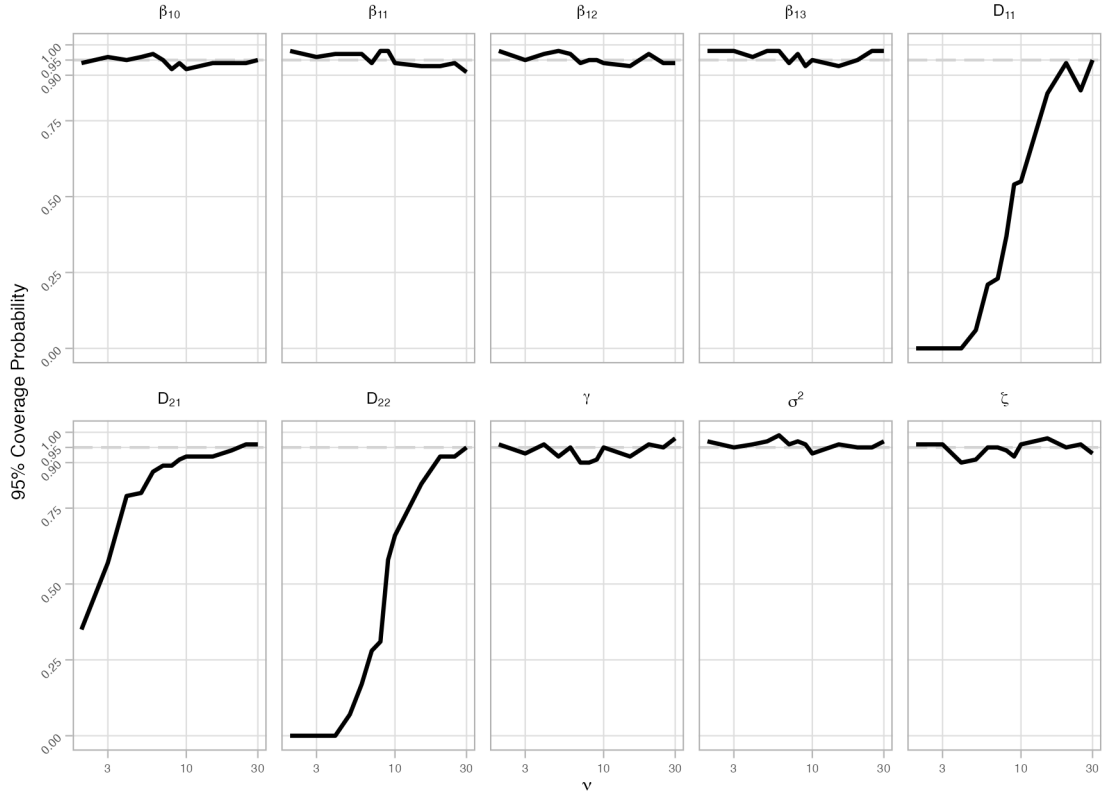


Figure 3.8: 95% Coverage for univariate simulation scenario with true values of random effects simulated from a bivariate t_ν distribution. A $\log_{10}$ scale is employed on the x-axis between ticks. The nominal value 0.95 is represented by a dashed line.

then expect this to ameliorate as $\nu \to \infty$ i.e. the true distribution approaches the normal.

Carrying out a brief simulation study – acting as a sensitivity analysis – allows inspection of exactly how poor said estimation of vech(D) is for smaller ν; whether the rest of the parameters $\boldsymbol{\Omega}_{-\mathrm{D}}$ suffer to any degree as well as how 'normal' the true distribution of the random effects, as determined by ν, need be for the approximation (3.1) to be reasonable.

We simulate a univariate joint model, with fixed effects with $\mathrm{vech}(\mathrm{D}) = (0.25, 0.00, 0.05)$. For each candidate $\nu \in \{2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30\}$ we simulate one hundred datasets, each with $n = 250$ and $r = 10$. The maximal degrees of freedom considered, $\nu = 30$, is where the t-distribution fully approximates the normal.

The results for the univariate joint models fit for this sensitivity analysis are tabulated in Table 3.1 for a subset of candidate ν values. Investigating first the parameter estimates $\hat{\boldsymbol{\Omega}}_{-\hat{\mathrm{D}}}$, we note there is little to distinguish between results across candidate degrees of freedom. The capabilities in estimation of $\mathrm{vech}\big(\hat{\mathrm{D}}\big)$ are, as we expected, very poor for smaller degrees of freedom; this improves as ν increases (i.e. the underlying distribution

Parameter	Mean (SD)	SE	Bias	MSE	CP	Mean (SD)	SE	Bias	MSE	CP
	$\nu = 2^*$					$\nu = 10$				
$D_{11} = 0.250$	2.098 (1.812)	0.089	1.848	6.667	0.00	0.311 (0.041)	0.032	0.061	0.005	0.55
$D_{21} = 0.000$	0.027 (0.566)	0.044	0.027	0.318	0.35	0.000 (0.013)	0.011	0.000	0.000	0.92
$D_{22} = 0.050$	0.340 (0.431)	0.017	0.290	0.268	0.00	0.061 (0.009)	0.007	0.011	0.000	0.66
$\beta_{10} = 2.000$	1.989 (0.150)	0.157	-0.011	0.022	0.94	1.993 (0.061)	0.056	-0.007	0.004	0.92
$\beta_{11} = 0.330$	0.314 (0.034)	0.041	-0.016	0.001	0.98	0.325 (0.021)	0.019	-0.005	0.000	0.94
$\beta_{12} = -0.500$	-0.489 (0.085)	0.098	0.011	0.007	0.98	-0.499 (0.042)	0.040	0.001	0.002	0.94
$\beta_{13} = 0.250$	0.285 (0.203)	0.224	0.035	0.042	0.98	0.260 (0.074)	0.080	0.010	0.006	0.95
$\sigma^2 = 0.160$	0.161 (0.006)	0.006	0.001	0.000	0.97	0.160 (0.006)	0.006	0.000	0.000	0.93
$\gamma = 0.500$	0.516 (0.074)	0.081	0.016	0.006	0.96	0.499 (0.144)	0.140	-0.001	0.021	0.95
$\zeta = -0.200$	-0.166 (0.285)	0.273	0.034	0.081	0.96	-0.189 (0.247)	0.251	0.011	0.060	0.96
	$\nu = 3$					$\nu = 20$				
$D_{11} = 0.250$	0.731 (0.492)	0.045	0.481	0.471	0.00	0.268 (0.032)	0.030	0.018	0.001	0.94
$D_{21} = 0.000$	0.006 (0.097)	0.019	0.006	0.009	0.57	0.002 (0.010)	0.010	0.002	0.000	0.94
$D_{22} = 0.050$	0.135 (0.050)	0.010	0.085	0.010	0.00	0.055 (0.006)	0.006	0.005	0.000	0.92
$\beta_{10} = 2.000$	2.009 (0.082)	0.089	0.009	0.007	0.96	2.003 (0.057)	0.054	0.003	0.003	0.94
$\beta_{11} = 0.330$	0.320 (0.025)	0.028	-0.010	0.001	0.96	0.326 (0.019)	0.018	-0.004	0.000	0.93
$\beta_{12} = -0.500$	-0.491 (0.059)	0.062	0.009	0.004	0.95	-0.498 (0.037)	0.038	0.002	0.001	0.97
$\beta_{13} = 0.250$	0.241 (0.117)	0.125	-0.009	0.014	0.98	0.239 (0.078)	0.075	-0.011	0.006	0.95
$\sigma^2 = 0.160$	0.160 (0.006)	0.006	0.000	0.000	0.95	0.162 (0.006)	0.006	0.002	0.000	0.95
$\gamma = 0.500$	0.517 (0.099)	0.099	0.017	0.010	0.93	0.519 (0.149)	0.147	0.019	0.022	0.96
$\zeta = -0.200$	-0.236 (0.244)	0.254	-0.036	0.060	0.96	-0.208 (0.248)	0.249	-0.008	0.061	0.95
	$\nu = 5$					$\nu = 30$				
$D_{11} = 0.250$	0.403 (0.063)	0.036	0.153	0.027	0.06	0.267 (0.032)	0.031	0.017	0.001	0.95
$D_{21} = 0.000$	0.001 (0.021)	0.013	0.001	0.000	0.80	0.001 (0.010)	0.010	0.001	0.000	0.96
$D_{22} = 0.050$	0.082 (0.017)	0.007	0.032	0.001	0.07	0.053 (0.007)	0.006	0.003	0.000	0.95
$\beta_{10} = 2.000$	2.007 (0.062)	0.064	0.007	0.004	0.96	1.993 (0.051)	0.053	-0.007	0.003	0.95
$\beta_{11} = 0.330$	0.325 (0.020)	0.021	-0.005	0.000	0.97	0.327 (0.019)	0.017	-0.003	0.000	0.91
$\beta_{12} = -0.500$	-0.499 (0.043)	0.045	0.001	0.002	0.98	-0.495 (0.039)	0.038	0.005	0.002	0.94
$\beta_{13} = 0.250$	0.242 (0.084)	0.091	-0.008	0.007	0.98	0.248 (0.070)	0.075	-0.002	0.005	0.98
$\sigma^2 = 0.160$	0.160 (0.006)	0.006	0.000	0.000	0.97	0.160 (0.005)	0.006	0.000	0.000	0.97
$\gamma = 0.500$	0.534 (0.121)	0.122	0.034	0.016	0.92	0.489 (0.151)	0.152	-0.011	0.023	0.98
$\zeta = -0.200$	-0.195 (0.260)	0.253	0.005	0.067	0.91	-0.215 (0.250)	0.252	-0.015	0.062	0.93

Table 3.1: Results from univariate simulation scenario with true values of random effects simulated from a bivariate t_ν distribution. 'Mean (SD)' denotes the average value (SE) from the one hundred model fits (i.e. an empirical summary) and 'SE' the average estimated standard error. 'MSE': Mean squared error; 'CP': 95% Coverage Probability. *: One model failed to converge for $\nu = 2$.

of the random effects approaches the normal). The 'progression' of coverage obtained for each ν is presented in Figure 3.8, wherein our interest is perhaps focused on elements of $\text{vech}\big(\hat{\text{D}}\big)$. We note, as expected, that as $\nu \to 30$ (i.e. approaches the normal) coverage of these parameters improves, achieving nominal 95% coverage. The remaining parameters $\hat{\boldsymbol{\Omega}}_{-\hat{\text{D}}}$ appear to be largely unaffected as ν increases and 'bounce' around the nominal 0.95 coverage line; perhaps most interestingly here is that the MLEs $\hat{\boldsymbol{\gamma}}$ appear robust to the misspecification of random effects variance-covariance at lower ν.

We argue, to some extent, that the covariance matrix D is often only of secondary interest (if indeed of *any* interest at all); the performative capabilities of $\boldsymbol{\Omega}_{-\text{D}}$ under the heavier-tailed distributions is perhaps perhaps satisfactory in this regard. However, although model performance does not deteriorate from a parameter estimation perspective as outlined, McFetridge et al. (2021) detail that misspecification of the random effects can lead to problems and inefficiencies with other aspects of the fitted joint model, such as individualised predictions, which we go on to outline in Chapter 6.

3.6.2 Censoring rate Υ

An increased number of censored observations during follow-up can introduce bias in the parameter estimates and widen the uncertainty in the parameter estimates, which is largely attributable to reduced information being available for estimating the hazard. In 'real' data analysis, a higher censoring rate might be due to some extraneous factor such as drop out due to a treatment side effect, and the Cox PH may subsequently incorrectly attribute the observed differences to the treatment itself, leading to confused/misleading inference which actually reflects these dropout-related characteristics rather than the risk of the event of interest.

Here, given the consternation surrounding the effect a greater number of censored observations may have on the Cox PH model itself, we seek to elucidate any sensitivity the joint model – particularly one fit by the approximate EM algorithm – may have. Additionally, one could posit since the contribution of the the survival density $\log f(T_i, \Delta_i|\boldsymbol{b}_i; \boldsymbol{\Omega})$ to the complete data log-likelihood under an increased censoring rate will be much smaller, the approximation in Section 3.2.1 may falter slightly. However, as pointed out by Rizopoulos (2012a), the complete data log-likelihood is *already* dominated by the contribution from the linear mixed models $\log f(\boldsymbol{Y}_i|\boldsymbol{b}_i; \boldsymbol{\Omega})$; we then anticipate that the approximation withstands this change to the underlying likelihood.

In our simulation studies as outlined in Sections 2.5.3 and 3.4.1 we control the rate of censoring by the rate parameter Υ. In the simulation studies carried out in Section 3.4, the censoring rate $\Upsilon = e^{-3.5}$ results in approximately 13% of subjects being censored during follow-up i.e. $T_i = C_i < \kappa$. In addition to this 'baseline' amount we consider two additional Υ with resultant increased censoring rates

$$\begin{aligned} \Upsilon = e^{-2.6} &\longrightarrow \approx 30\% \text{ censoring;} \\ \Upsilon = e^{-1.9} &\longrightarrow \approx 50\% \text{ censoring.} \end{aligned}$$

The estimates for $\hat{\boldsymbol{\Phi}}$ are presented in Figure 3.9, here we note that the spread of estimates appear to increase with the amount of censoring taking place. An in-depth tabulation is presented in Appendix B.1.6 wherein we observe little to distinguish between $\hat{\boldsymbol{\Omega}}_{-\hat{\boldsymbol{\Phi}}}$ save for parameter estimates becoming more conservative on average as the censoring rate increases.

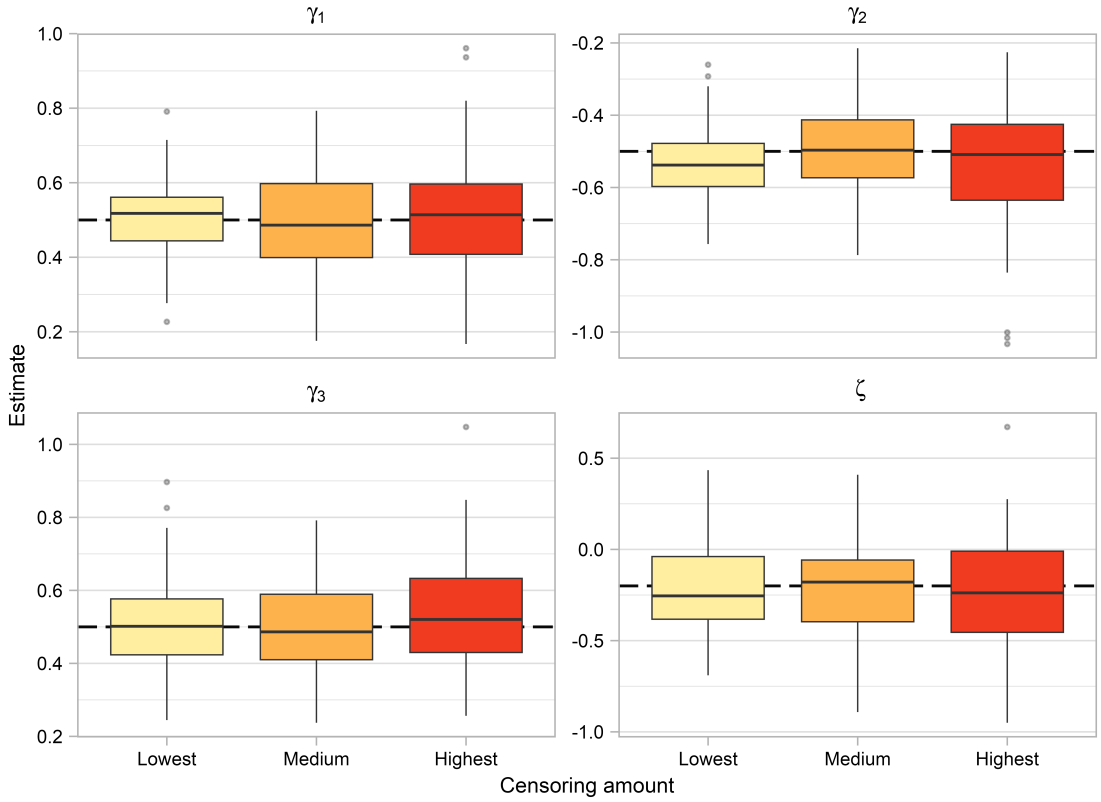


Figure 3.9: Estimates for survival parameters $\boldsymbol{\gamma}$ and ζ for different censoring rates. 'Lowest' denotes $\Upsilon = e^{-3.5}$; 'Medium' denotes $\Upsilon = e^{-2.6}$; and 'Highest' $\Upsilon = e^{-1.9}$. The dashed line signifies the true parameter value.

3.7 Note on implementation

Since faster computation is at the 'heart' of the research objective, we bring this chapter to a close by briefly bridging from the theory (in both this chapter and the previous one), results in Sections 3.2 and 3.3, with practical details. Here, we seek to offer some insight into underpinning computational details.

The statistical programming language `R` (R Core Team, 2020) is utilised throughout, with computational bottlenecks rewritten in `C++`, facilitated by `R` package `Rcpp` (Eddelbuettel and François, 2011), which provides seamless integration between the R programming language and `C++`: Allowing `R` users to write high-performance, computationally intensive functions in `C++` and easily call them from within `R`. In addition to the many 'standard' `C` libraries integrated via `Rcpp`, we make use of an additional package `RcppArmadillo` (Eddelbuettel and Sanderson, 2014), which further integrates the Armadillo `C++` library with `R`. Armadillo is a popular library, which is used for linear algebra (which we carry out a lot of), and balances intuitive, `MATLAB`-esque syntax with high performance implementations

of many numerical operations. In the `optim` step in Section 3.2.4 a great deal of computational efficiency can be gained by providing the gradient vector $\frac{\partial \log f(\boldsymbol{b}_i, T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega})}{\partial \boldsymbol{b}_i}$, whose form is given in Appendix A.5.

An example of the gains in performance for evaluating the conditional expectation (3.14), which represents a significant computational bottleneck in practise, is provided in Appendix D.

Chapter 4

Faster Fitting: Towards Fast & Flexible Joint Models

The work presented in this chapter is based on the publication:
Murray, J., Philipson, P., 2023. *Fast estimation for generalised multivariate joint models using an approximate EM algorithm.* Computational Statistics & Data Analysis 187, 107819. doi: 10.1016/j.csda.2023.107819.

4.1 Generalised linear mixed models

Hitherto, we have solely considered joint models under the 'classic' framework of longitudinal outcomes and an event-time process in Chapters 2 and 3 which are ubiquitous in literature. The key assumption therein being that the (typically continuous) longitudinal responses are Gaussian, or their transformations. However, this assumption may be unrealistic in clinical settings, where transformations may complicate interpretation.

We consider an alternative modelling approach: Seeking to remove the Gaussian restriction on the model for the longitudinal process (2.1). In this section, we vacate the joint modelling landscape habituated up to this point as we introduce such an extension; namely generalised linear mixed models (GLMMs).

4.1.1 Formulation of GLMMs

Instead of the LMM presented in (2.1), we now model the expected value of the response conditioned on the random effects

$$h(\mathbb{E}[\boldsymbol{Y}_i|\boldsymbol{b}_i]) = \boldsymbol{\eta}_i = \mathrm{X}_i\boldsymbol{\beta} + \mathrm{Z}_i\boldsymbol{b}_i, \tag{4.1}$$

where $h(\cdot)$ is a known, monotonic link function providing the relationship between the conditional distribution of the response $\boldsymbol{Y}_i$ and the linear predictor $\boldsymbol{\eta}_i$. In (4.1), X_i and Z_i share the same definition as previously seen in (2.1) in Section 2.1.1. We note that the LMM (2.1) is captured by the GLMM framework (4.1) with $h(\cdot)$ set as the identity function.

To estimate parameters and make inference under a GLMM we require the marginal likelihood $f(\boldsymbol{Y}_i|\boldsymbol{\Omega})$, since it represents the probability of observing $\boldsymbol{Y}_i$ given parameters $\boldsymbol{\Omega}$. Furthermore, this quantity *allows* us to proceed by maximum likelihood (i.e. since we iteratively find the set of parameters which maximise it), or indeed by Bayesian methods since the marginal likelihood is present in the posterior sampled. The log-likelihood (marginal over the random effects) can be written as

$$\ell(\boldsymbol{\Omega}) = \sum_{i=1}^{n} \log f(\boldsymbol{Y}_i|\boldsymbol{\Omega}) = \log \int f(\boldsymbol{Y}_i|\boldsymbol{b}_i;\boldsymbol{\Omega})\, f(\boldsymbol{b}_i|\boldsymbol{\Omega})\, d\boldsymbol{b}_i. \tag{4.2}$$

For explicitness' sake, in the above, $f(\boldsymbol{Y}_i|\boldsymbol{b}_i;\boldsymbol{\Omega})$ is the conditional likelihood of the response given the random effects and parameters $\boldsymbol{\Omega}$, and is 'decided' based on the specification of the model (i.e. the conditional distribution $\boldsymbol{Y}_i|\boldsymbol{b}_i$); the next term $f(\boldsymbol{b}_i|\boldsymbol{\Omega})$ is the conditional likelihood of the random effects, which as previously circumscribed in (2.1) are *always* multivariate normal. Overall, the right hand side of (4.2) simply averages the response

over all values of the random effects given the model parameters.

Owing to this integral, no closed form solution exists, instead requiring the use of numerical integration techniques such as those surveyed (albeit under the setting of joint models) in Sections 2.3.2–2.3.5.

4.1.2 Advantages of GLMMs

Modelling a response of interest by a GLMM instead of employing an LMM on a transformed version of said response is likely advantageous. As we previously outlined in Section 1.2.2, a biomarker of interest may come in the form of a binary presence/absence; questionnaire scores, for instance, will manifest as counts; skewed continuous data e.g. waiting times will most likely not be Gaussian; and Likert-type responses are frequently used in clinical trials e.g. pain scores. In these cases, the practitioner can consider the 'most natural' distribution for the data at hand.

In addition to the greater breadth of response distributions one can consider for a response, GLMMs for count models can be extended to accommodate an undue excess of zeroes in the data (zero-inflation), or structural absence of zeroes (zero-truncation). Furthermore, the dispersion defined as the ratio of variance to the mean of the response can also be modelled, potentially improving the model's goodness of fit. GLMMs therefore provide a flexible modelling framework which arms the statistician with a great many 'levers to pull'.

However, whilst there are numerous advantages to this modelling route, there is an accompanying uptake in computational demand. The introduction of many model extensions in addition to the already-present random effects specification may lead to more frequent model over-fitting[1]. Indeed, one may feel overwhelmed by choice; choosing a complex, computationally expensive model which does not significantly outperform a simpler one and so on.

[1]In the context of a GLMM, over-fitting usually occurs due to fitting too many random effects, which in actuality capture noise in the data, the model then attributing this to purported variance amongst the random effects, such that the resultant model is not generalisable. Given then the additional inclusion of a dispersion model which alters the variance structure, over-fitting may occur more readily owing to the presence of less variation given the dispersion model, which the model still attempts to attribute to the random effects

4.2 Beyond Gaussian assumptions and the Bayesian paradigm in joint modelling

With GLMMs established in Section 4.1, we turn attention now to their inclusion in a joint model. We note the relatively rapid emergence of joint models with (at least one) GLMM sub-model in Section 1.2.2, given the onus of accommodating diverse response types.

We now proceed by describing the exclusive use of the Bayesian paradigm for fitting these 'generalised' joint models, before re-introducing the approximation used in Chapter 3 to return to the maximum likelihood framework utilised throughout.

4.2.1 Beyond the Gaussian assumption: Accommodating diverse longitudinal outcomes

The assumption that the longitudinal response(s) of interest are Gaussian may be for convenience of implementation: It allows for us to consider run-of-the-mill software to fit LMMs, as well as allowing for closed-form updates to the fixed effect and dispersion parameters we outlined in (2.11), and shown under the approximate EM algorithm in Sections 3.3.2 and 3.3.3, which compound said relative computational ease.

Clearly then, there exists something of a 'market' to be able to support joint models whose longitudinal sub-models are of differing types; with burgeoning examples in literature – as well as software availability – surrounding them. Advantages of this modelling approach was outlined in Section 4.1.2. We note that these *exclusively exist* in the Bayesian framework, to the best of the author's knowledge.

4.2.2 Reliance on Bayesian approaches

Hitherto, implementations of GLMM sub-models in a multivariate joint model setting have been predicated on the aforementioned Bayesian paradigm – to the best of the author's knowledge – with no implementation via maximum likelihood, which is closer in spirit to well-established methods for analysing clinical data such as the Cox model and linear mixed effects model for survival and longitudinal data respectively; with disquisition given in Chapter 2 in addition to classic treatise by Tsiatis and Davidian (2004). If MVJMs are to find their way into routine clinical use in a manner akin to Cox PH and LMMs it is important to at least have the *option* of a non-Bayesian implementation; the work presented in this Chapter attempts to fill this current void.

Ostensibly, up to this point in joint modelling literature, only MCMC (or indeed, approximate Bayesian inference in the case of `INLA`) can accommodate the joint models

described in the previous section. Said reliance on MCMC methods harkens back to the 'expanded form' taken by the conditional expectations required by the semiparametric maximum likelihood approach undertaken in Sections 2.2 and 2.3. Namely, we noted that the conditional density $f(\boldsymbol{b}_i|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega})$ takes the form (2.16)

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

Recall additionally, we previously enjoyed – under the assumed (multivariate) normality of the response $\boldsymbol{Y}_i|\boldsymbol{b}_i$ – a *tractable* expression for the density $f(\boldsymbol{b}_i|\boldsymbol{Y}_i; \boldsymbol{\Omega})$ given by multivariate normal theory (2.18), which acted as a foundation for *nearly all* maximum likelihood implementations of the 'classic' joint model, since it allowed for evaluation of expectations against the target density (2.13) whose expression existed in closed-form via e.g. quadrature methods (Wulfsohn and Tsiatis, 1997), as described in Sections 2.3.1–2.3.4.

Issues arise under maximum likelihood then in circumstances where the density $f(\boldsymbol{b}_i|\boldsymbol{Y}_i; \boldsymbol{\Omega})$ *doesn't* enjoy a neat, tractable expression. Instead, one must evaluate the density (2.15), wherein the marginal $f(\boldsymbol{Y}_i|\boldsymbol{\Omega})$ involves integration over the random effects; at best some workable expression may well exist for certain exponential families, but perhaps owing to the wider use of the multivariate normal, they remain comparatively buried.

4.2.3 Maximum likelihood approach via the approximate EM algorithm

In Chapter 3 we demonstrated use of an approximation on the conditional distribution of the random effects (Bernhardt et al., 2015; Murray and Philipson, 2022). This approximation (3.1) effectively 'collapses' down the *entire* conditional density $f(\boldsymbol{b}_i|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega})$ to the approximate normal distribution $N\big(\hat{\boldsymbol{b}}_i, \hat{\Sigma}_i\big)$, with $\hat{\boldsymbol{b}}_i$ and $\hat{\Sigma}_i$ given in (3.2) and (3.3), respectively. Use of this approximation inherently then eschews *any* consideration of the form of $f(\boldsymbol{b}_i|\boldsymbol{Y}_i; \boldsymbol{\Omega})$, the density causing consternation around any possible non-Gaussian response $\boldsymbol{Y}_i|\boldsymbol{b}_i$ in the previous section, instead allowing us to consider $f(\boldsymbol{b}_i|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega})$ by the same approximation, with a non-Gaussian model for $\mathbb{E}_i[\boldsymbol{Y}_i|\boldsymbol{b}_i]$.

Use of the approximation (3.1) here, since we considered GLMMs at large, additionally then calls on results outside of joint modelling, where the random effects for *any* GLMM are shown to be asymptotically normal (Baghishani and Mohammadzadeh, 2012), something we appeal to here as we consider diverse longitudinal responses whose constituent submodel is defined by (4.1).

4.3 Generalised multivariate joint models

4.3.1 A flexible longitudinal specification

Revisiting the formulation laid-out in Section 2.1, we now consider the case where we assume the conditional distribution of the k^{th} response belongs to a member of the exponential family, such that it is modelled by the GLMM (4.1), introduced in Section 4.1.1. The linear mixed model in (2.1) is replaced by a GLMM for each $\boldsymbol{Y}_{ik}$ with linear predictor $\boldsymbol{\eta}_{ik}$

$$h_k(\mathbb{E}_i[\boldsymbol{Y}_{ik}|\boldsymbol{b}_{ik}]) = \boldsymbol{\eta}_{ik} = \mathrm{X}_{ik}\boldsymbol{\beta}_k + \mathrm{Z}_{ik}\boldsymbol{b}_{ik}, \tag{4.3}$$

where $h_k(\cdot)$ denotes the known monotonic link function imposed on the k^{th} response. We then form the joint model previously outlined in Section 2.1.1 by inducing an association between the random effects present in linear predictor $\boldsymbol{\eta}_{ik}$ and the hazard $\lambda_i(t)$ as previously defined in (2.2). The observed likelihood function for subject i is altered slightly to

$$\begin{aligned} f(T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega}) &= \int_{-\infty}^{\infty} f(T_i, \Delta_i, \boldsymbol{Y}_i, \boldsymbol{b}_i; \boldsymbol{\Omega})\, d\boldsymbol{b}_i \\ &= \int_{-\infty}^{\infty} \left[\prod_{k=1}^{K} f(\boldsymbol{Y}_{ik}|\boldsymbol{b}_{ik}; \boldsymbol{\beta}_k, \boldsymbol{\sigma}_k) \right] f(T_i, \Delta_i|\boldsymbol{b}_i; \boldsymbol{\gamma}, \boldsymbol{\zeta})\, f(\boldsymbol{b}_i|\mathrm{D})\, d\boldsymbol{b}_i, \end{aligned} \tag{4.4}$$

wherein $f(\boldsymbol{Y}_{ik}|\cdot)$ now corresponds to an appropriate probability density function for the k^{th} longitudinal response with dispersion parameters (if applicable) $\boldsymbol{\sigma}_k$. We explicitly note that the survival sub-model (2.2) remains unchanged.

As was the case under the 'classic' setting in Section 2.1.2, we once more collate random effects $\boldsymbol{b}_i = \left(\boldsymbol{b}_{i1}{}^\top, \ldots, \boldsymbol{b}_{iK}{}^\top\right)^\top$ for subject i; fixed effects $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_K^\top\right)^\top$; association parameters $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)^\top$; and now dispersion parameters $\boldsymbol{\sigma} = \left(\boldsymbol{\sigma}_1^\top, \ldots, \boldsymbol{\sigma}_K^\top\right)^\top$. If the k^{th} sub-model does *not* have a dispersion model (which we describe in the next section), its element is simply set to zero and ignored. The vector notation used for the dispersion parameter $\boldsymbol{\sigma}_k$ is also the simply general case; for e.g. the Gaussian we considered in Chapters 2 and 3, this quantity simply represents (lazily) the scalar residual variance $\boldsymbol{\sigma} = \sigma_\varepsilon^2$.

We then construct the parameter vector $\boldsymbol{\Omega} = \left(\mathrm{vech}(\mathrm{D})^\top, \boldsymbol{\beta}^\top, \boldsymbol{\sigma}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{\zeta}^\top\right)^\top$ for what we term the "generalised" multivariate joint model (4.4) ('GMVJM').

4.3.2 Dispersion models

As alluded to in the previous section, if $\boldsymbol{Y}_{ik}|\boldsymbol{b}_{ik}$ is assumed to be distributed by (and subsequently modelled with) an exponential family member with dispersion parameter(s) $\boldsymbol{\sigma}_k$, then it is pertinent to be able to appropriately model this dispersion 'within' the k^{th} longitudinal sub-model. That is, allowing for the practitioner to impose beliefs on the nature of this dispersion as part of the joint model; doing so may affect the random effects $\boldsymbol{b}_{ik}$, thereby affecting $\boldsymbol{\gamma}$ and possibly inference as a whole. Note we explicitly state which distributions we consider in the upcoming Section 4.3.3, whilst in this short section we simply describe this dispersion model formulation. We introduce

$$\tilde{h}_k(\boldsymbol{\varphi}_{ik}) = \tilde{\boldsymbol{\eta}}_{ik} = \mathrm{W}_{ik}\boldsymbol{\sigma}_k, \tag{4.5}$$

where W_{ik} is the design matrix associated with the k^{th} response for subject i, and $\tilde{h}_k(\cdot)$ is the known, monotonic, link function imposed on this linear predictor for the dispersion model for the k^{th} longitudinal sub-model, $\tilde{\boldsymbol{\eta}}_{ik}$.

We note a few special cases. In circumstances where dispersion is defined by an 'intercept only' specification, W_{ik} takes the form of a column vector of ones such that $\mathrm{W}_{ik}\boldsymbol{\sigma}_k$ resolves to a scalar quantity; if there is no dispersion associated with the model then an empty matrix is used (since the corresponding $\sigma_k \in \boldsymbol{\sigma}$ is zero and ignored); in the Gaussian case, W_{ik} is set to one, such that one obtains the residual variance $\sigma^2_{\varepsilon_k}$.

We note that the notation used here is a little confusing, mostly in its use across all GLMMs, but we believe it serves as a neat 'catch-all', with its special cases written above.

4.3.3 Considered distributions for the longitudinal response

With the formulation of the longitudinal process, and optional dispersion models laid out in Sections 4.3.1 and 4.3.2, respectively, we now state candidate exponential families for the longitudinal response. Table 4.1 provides an overview for these candidate distributions, with following small sections seeking to provide more detail. Note in both Table 4.1 and the subsequent sections, we eschew the k subscript for notational convenience.

Remark. In log-likelihoods (and following calculations upon them), the element-wise product of two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is written simply as $\boldsymbol{xy}$ instead of the perhaps more correct $\boldsymbol{x} \odot \boldsymbol{y}$. This is simply due to formatting/space constraints (and secondarily due to the author's personal preference).

Distribution	Link	Domain	$f(\boldsymbol{Y}_i\|\boldsymbol{b}_i;\ \cdot)$
$\boldsymbol{Y}_i\|\boldsymbol{b}_i, \sigma^2 \sim N(\boldsymbol{\mu}_i, \mathrm{V}_i)$	h: $\mathbb{E}[\boldsymbol{Y}_i]$	$\boldsymbol{Y}_i \in \mathbb{R}$; $\boldsymbol{\mu}_i \in \mathbb{R}$	$-\frac{m_i}{2}\log 2\pi - \frac{1}{2}\log\lvert\mathrm{V}_i\rvert$ $-\frac{1}{2}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)^\top \mathrm{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)$
$\boldsymbol{Y}_i\|\boldsymbol{b}_i \sim \mathrm{Po}(\boldsymbol{\mu}_i)$	h: $\log(\mathbb{E}[\boldsymbol{Y}_i])$; $\tilde{h}$: N/A	$\boldsymbol{Y}_i \in \mathbb{N}_0$; $\boldsymbol{\mu}_i \in \mathbb{R}^+$	$\exp\{\boldsymbol{\mu}_i\}\boldsymbol{\mu}_i^{\boldsymbol{Y}_i}/\boldsymbol{Y}_i!$
$\boldsymbol{Y}_i\|\boldsymbol{b}_i \sim \mathrm{Bin}(\boldsymbol{\mu}_i)$	h: $\mathrm{logit}(\mathbb{E}[\boldsymbol{Y}_i])$; $\tilde{h}$: N/A	$\boldsymbol{Y}_i \in \{0, 1\}$; $\boldsymbol{\mu}_i \in [0, 1]$	$\boldsymbol{\mu}_i^{\boldsymbol{Y}_i}(\mathbf{1} - \boldsymbol{\mu}_i)^{(\mathbf{1}-\boldsymbol{Y}_i)}$
$\boldsymbol{Y}_i\|\boldsymbol{b}_i; \boldsymbol{\sigma} \sim \mathrm{NB}(\boldsymbol{\mu}_i, \boldsymbol{\varphi}_i)$	h: $\log(\mathbb{E}[\boldsymbol{Y}_i])$; $\tilde{h}$: $\log(\boldsymbol{\varphi}_i)$	$\boldsymbol{Y}_i \in \mathbb{N}_0$; $\boldsymbol{\mu}_i \in \mathbb{R}^+$	$\frac{\Gamma(\boldsymbol{Y}_i+\boldsymbol{\varphi}_i)\boldsymbol{\varphi}_i^{\boldsymbol{\varphi}_i}\boldsymbol{\mu}_i^{\boldsymbol{Y}_i}}{\Gamma(\boldsymbol{\varphi}_i)\Gamma(\boldsymbol{Y}_i+1)(\boldsymbol{\mu}_i+\boldsymbol{\varphi}_i)^{\boldsymbol{\varphi}_i+\boldsymbol{Y}_i}}$
$\boldsymbol{Y}_i\|\boldsymbol{b}_i; \boldsymbol{\sigma} \sim \mathrm{GP}(\boldsymbol{\mu}_i, \boldsymbol{\varphi}_i)$	h: $\log(\mathbb{E}[\boldsymbol{Y}_i])$; $\tilde{h}$: $\boldsymbol{\varphi}_i$	$\boldsymbol{Y}_i \in \mathbb{N}_0$; $\boldsymbol{\mu}_i \in \mathbb{R}^+$	$\frac{\boldsymbol{\mu}_i(\boldsymbol{\mu}_i+\boldsymbol{Y}_i\boldsymbol{\varphi}_i)^{\boldsymbol{Y}_i-1}}{(1+\boldsymbol{\varphi}_i)^{\boldsymbol{Y}_i}\boldsymbol{Y}_i!}\exp\left\{-\frac{\boldsymbol{\mu}_i+\boldsymbol{\varphi}_i\boldsymbol{Y}_i}{1+\boldsymbol{\varphi}_i}\right\}$
$\boldsymbol{Y}_i\|\boldsymbol{b}_i; \boldsymbol{\sigma} \sim \mathrm{Ga}(\boldsymbol{\vartheta}_i, \boldsymbol{\varphi}_i)$ $\boldsymbol{\vartheta}_i = \boldsymbol{\mu}_i/\boldsymbol{\varphi}_i$	h: $\log(\mathbb{E}[\boldsymbol{Y}_i])$; $\tilde{h}$: $\log(\boldsymbol{\varphi}_i)$	$\boldsymbol{Y}_i \in \mathbb{R}^+$; $\boldsymbol{\mu}_i \in \mathbb{R}^+$	$\frac{\boldsymbol{Y}_i^{\boldsymbol{\varphi}_i-1}}{\Gamma(\boldsymbol{\varphi}_i)\boldsymbol{\vartheta}_i^{\boldsymbol{\vartheta}_i}}\exp\left\{-\frac{\boldsymbol{Y}_i}{\boldsymbol{\vartheta}_i}\right\}$

Table 4.1: Candidate exponential distributions conditioned on the random effects; the link functions as used in (4.3) and (4.5); the domain occupied by the response $\boldsymbol{Y}_i$ and the mean $\boldsymbol{\mu}_i = h^{-1}(\boldsymbol{\eta}_i)$; and the PMF/PDF. 'Po': Poisson; 'Bin': Binomial; 'NB': Negative binomial; 'GP': Generalised Poisson; 'Ga': Gamma. Explanation for terms in the Gaussian case are given in Section 2.1.2. $\mathbb{N}_0$ denotes the set of natural numbers with zero included. $\Gamma(\cdot)$ is the Gamma function. The linear predictor is invariably defined as $\boldsymbol{\eta}_i = \mathrm{X}_i\boldsymbol{\beta} + \mathrm{Z}_i\boldsymbol{b}_i$.

Poisson

The most obvious distribution for count data. The expectation and variance of the (conditionally) Poisson distributed response is given by $\mathbb{E}[\boldsymbol{Y}_i] = \mathrm{Var}[\boldsymbol{Y}_i] = \boldsymbol{\mu}_i$. No dispersion is considered in the Poisson case. The log-likelihood contribution from subject i, ℓ_i, under the Poisson GLMM is

$$\ell_i = \log f(\boldsymbol{Y}_i|\boldsymbol{b}_i; \boldsymbol{\Omega}) = \boldsymbol{Y}_i^\top \log \boldsymbol{\mu}_i - \mathbf{1}^\top \boldsymbol{\mu}_i - \mathbf{1}^\top \log \boldsymbol{Y}_i! \tag{4.6}$$

here $\mathbf{1}$ represents an m_i-vector of ones, such that e.g. $\mathbf{1}^\top\boldsymbol{x}$ resolves to a scalar containing the sum of vector $\boldsymbol{x}$. The above assumption of equal mean and variance (*equidispersion*) is, in actuality, a strong assumption. In fact, some classic texts admonish that equidispersion is the *exception* in practise, and quite often violated (McCullagh and Nelder, 1989).

Binomial

The sole distribution considered for a binary response which takes values 0 or 1, with mean $\boldsymbol{\mu}_i = \exp\{\boldsymbol{\eta}_i\}/(\mathbf{1} + \exp\{\boldsymbol{\eta}_i\}) \in [0, 1]$. The expectation of the binomially distributed response is $\mathbb{E}[\boldsymbol{Y}_i] = \boldsymbol{\mu}_i$ (i.e. the probability) with variance $\mathrm{Var}[\boldsymbol{Y}_i] = \boldsymbol{\mu}_i(\mathbf{1} - \boldsymbol{\mu}_i)$. No

dispersion is modelled. The log-likelihood of the binomial GLMM is

$$\ell_i = \log f(\boldsymbol{Y}_i|\boldsymbol{b}_i;\boldsymbol{\Omega}) = \boldsymbol{Y}_i^\top \log \boldsymbol{\mu}_i + (\mathbf{1} - \boldsymbol{Y}_i)^\top \log(\mathbf{1} - \boldsymbol{\mu}_i)\,,$$

which we rewrite in terms of the linear predictor $\boldsymbol{\eta}_i$

$$\begin{aligned}\ell_i &= \boldsymbol{Y}_i^\top \log\left(\frac{\exp\{\boldsymbol{\eta}_i\}}{\mathbf{1}+\exp\{\boldsymbol{\eta}_i\}}\right) + (\mathbf{1} - \boldsymbol{Y}_i)^\top \log\left(\mathbf{1} - \frac{\exp\{\boldsymbol{\eta}_i\}}{\mathbf{1}+\exp\{\boldsymbol{\eta}_i\}}\right), \\ &= \boldsymbol{Y}_i^\top \log(\exp\{\boldsymbol{\eta}_i\}) - \boldsymbol{Y}_i^\top \log(\mathbf{1}+\exp\{\boldsymbol{\eta}_i\}) + (\mathbf{1} - \boldsymbol{Y}_i)^\top(-\log(\mathbf{1}+\exp\{\boldsymbol{\eta}_i\}))\,,\end{aligned}$$

which is equivalent to

$$\ell_i = \boldsymbol{Y}_i^\top \boldsymbol{\eta}_i - \mathbf{1}^\top \log(\mathbf{1}+\exp\{\boldsymbol{\eta}_i\})\,. \tag{4.7}$$

Negative binomial

As explained previously, the Poisson distribution's assumption of equal mean and variance is, in fact, quite a stringent one, which may be unrealistic in practise. If the variance is greater than the mean for some data, $\mathrm{Var}[\boldsymbol{Y}_i] > \mathbb{E}[\boldsymbol{Y}_i]$, then the data is *over*dispersed. Note that in the case of grouped (e.g. longitudinal) data, dispersion can also be measured within-group. Imposition of the Poisson in presence of overdispersion may underestimate standard errors of parameters; potentially prematurely declaring significance of its parameters.

A popular distribution for modelling overdispersed count data is the negative binomial. The log-likelihood for the negative binomial GLMM is

$$\begin{aligned}\ell_i = \log f(\boldsymbol{Y}_i|\boldsymbol{b}_i;\boldsymbol{\Omega}) =& \mathbf{1}^\top\{\log\Gamma(\boldsymbol{Y}_i+\boldsymbol{\varphi}_i) - \log\Gamma(\boldsymbol{\varphi}_i) - \log\Gamma(\boldsymbol{Y}_i+\mathbf{1})\} \\ &+ \boldsymbol{\varphi}_i^\top \log\boldsymbol{\varphi}_i - \boldsymbol{\varphi}_i^\top \log(\boldsymbol{\mu}_i+\boldsymbol{\varphi}_i) + \boldsymbol{Y}_i^\top[\log\boldsymbol{\mu}_i - \log(\boldsymbol{\mu}_i+\boldsymbol{\varphi}_i)]\,.\end{aligned} \tag{4.8}$$

In literature, the above parameterisation of the negative binomial is referred to as the Type 1 negative binomial ('NB-1', etc.). Under this parameterisation, the expected value of the response is $\mathbb{E}[\boldsymbol{Y}_i] = \boldsymbol{\mu}_i$ with variance $\mathrm{Var}[\boldsymbol{Y}_i] = \boldsymbol{\mu}_i + \boldsymbol{\mu}_i^2/\boldsymbol{\varphi}_i$.

For completeness' sake, we note that these NB models are sometimes (collectively) called Poisson-Gamma models, since they combine the Poisson distribution for modelling counts with a Gamma distribution, which accounts for the variability in the underlying rate parameter (Ismail and Jemain, 2007).

Generalised Poisson ('GP-1')

Although the previously seen negative binomial model accounts for overdispersion in the rate of a Poisson distribution, it is also possible for some given data to exhibit the oppo-

site, $\text{Var}[\boldsymbol{Y}_i] < \mathbb{E}[\boldsymbol{Y}_i]$, *under*dispersion. One popular approach for modelling *both* underdispersed *and* overdispersed count data is the generalised Poisson ('GP'). As with the NB approach, several parameterisations of the distribution exist. We opt for one introduced in Zamani and Ismail (2012), referred to therein as the 'GP-1' parameterisation. The log-likelihood of the GP-1 GLMM is

$$\begin{aligned}\ell_i = \log f(\boldsymbol{Y}_i|\boldsymbol{b}_i;\boldsymbol{\Omega}) =& \mathbf{1}^\top \log \boldsymbol{\mu}_i + (\boldsymbol{Y}_i - \mathbf{1})^\top \log(\boldsymbol{\mu}_i + \boldsymbol{\varphi}_i \boldsymbol{Y}_i) - \boldsymbol{Y}_i^\top \log(\mathbf{1} + \boldsymbol{\varphi}_i) \\ & - \mathbf{1}^\top \log \boldsymbol{Y}_i! - \mathbf{1}^\top \left\{ \frac{\boldsymbol{\mu}_i + \boldsymbol{\varphi}_i \boldsymbol{Y}_i}{\mathbf{1} + \boldsymbol{\varphi}_i} \right\}.\end{aligned} \tag{4.9}$$

The expectation of the GP-1 distributed response is simply the mean $\mathbb{E}[\boldsymbol{Y}_i] = \boldsymbol{\mu}_i$ with variance $\text{Var}[\boldsymbol{Y}_i] = (\mathbf{1} + \boldsymbol{\varphi}_i)^2 \boldsymbol{\mu}_i$. Zamani and Ismail (2012) point out that the dispersion parameter $\boldsymbol{\varphi}_i$ (the item we want to estimate) informs the dispersion *factor* $(\mathbf{1} + \boldsymbol{\varphi}_i)^2$, which tells us the multiplicative factor of difference between the mean and variance. Overdispersion occurs when $\boldsymbol{\varphi}_i > 0$ and underdispersion when $\boldsymbol{\varphi}_i < 0$; interestingly (4.9) reduces to the Poisson (4.6) when $\boldsymbol{\varphi}_i = \mathbf{0}$, trivial proof is provided in Appendix A.6. One drawback with this GP-1 model is that the dispersion parameter $\boldsymbol{\varphi}_i$ exists on a bounded scale

$$\max(-1, -\boldsymbol{\mu}_i/4) < \frac{\boldsymbol{\varphi}_i}{\mathbf{1} + \boldsymbol{\varphi}_i} < 1, \tag{4.10}$$

which could make estimation, as well as data generation, a tricky feat.

Gamma

The Gamma distribution is a worthy candidate for instances where the data is continuous, but imposing the normal distribution would be inappropriate, an example of this could be waiting times (or indeed anything time-related, as these distributions tend to be skew). The Gamma is parameterised by the shape $\boldsymbol{\varphi}_i \in \mathbb{R}^+$ and scale $\boldsymbol{\vartheta}_i = \boldsymbol{\mu}_i/\boldsymbol{\varphi}_i$ parameters. The log link is used instead of the canonical inverse link due to the latter option here being problematic (Bolker, 2022); the Gamma then being used when e.g. the log-normal distribution would also be a good candidate. The log-likelihood for the Gamma GLMM is

$$\begin{aligned}\ell_i = \log f(\boldsymbol{Y}_i|\boldsymbol{b}_i;\boldsymbol{\Omega}) =& (\boldsymbol{\varphi}_i - \mathbf{1})^\top \boldsymbol{Y}_i - \mathbf{1}^\top \left(\frac{\boldsymbol{Y}_i}{\boldsymbol{\vartheta}_i} \right) - \mathbf{1}^\top \log \Gamma(\boldsymbol{\varphi}_i) - \boldsymbol{\varphi}_i^\top \log \boldsymbol{\vartheta}_i \\ =& (\boldsymbol{\varphi}_i - \mathbf{1})^\top \boldsymbol{Y}_i - \mathbf{1}^\top \left(\frac{\boldsymbol{Y}_i \boldsymbol{\varphi}_i}{\boldsymbol{\mu}_i} \right) - \mathbf{1}^\top \log \Gamma(\boldsymbol{\varphi}_i) - \boldsymbol{\varphi}_i^\top \log \left(\frac{\boldsymbol{\mu}_i}{\boldsymbol{\varphi}_i} \right).\end{aligned} \tag{4.11}$$

The shape parameter $\boldsymbol{\varphi}_i$ determines (as the name suggests) the shape of the resulting Gamma distribution, encapsulating the resulting skewness (i.e. how 'quickly' or 'sharply' the pdf decays). By estimating this, we may better capture the characteristics of the

conditional distribution of the response, thereby better ascertaining its variability (over e.g. assuming log normality, as previously suggested).

4.4 Estimation for generalised multivariate joint models

With the observed data likelihood (4.4) established along with the candidate exponential family distributions we consider in this chapter, we can turn our attention toward estimation of the parameter vector $\boldsymbol{\Omega}$. As was described in Section 4.2.2, estimation of $\boldsymbol{\Omega}$ is exclusively done under a Bayesian framework in existing literature, with the following approach via maximum likelihood being a novel methodology to the best of the author's knowledge.

Following the necessary update to the observed data likelihood (4.3), we additionally need to slightly alter the E-step (2.10) to account for the K potentially different response families. At the EM step at iteration $(m+1)$ we seek to maximise

$$Q\Big(\boldsymbol{\Omega}; \boldsymbol{\Omega}^{(m)}\Big) = \sum_{i=1}^{n} \mathbb{E}_i \left[\left\{ \sum_{k=1}^{K} \log f\Big(\boldsymbol{Y}_{ik} | \boldsymbol{b}_{ik}; \boldsymbol{\Omega}^{(m)}\Big) \right\} + \log f\Big(T_i, \Delta_i | \boldsymbol{b}_i; \boldsymbol{\Omega}^{(m)}\Big) + \log f\Big(\boldsymbol{b}_i | \boldsymbol{\Omega}^{(m)}\Big) \right], \tag{4.12}$$

wherein the log densities for the survival process and random effects are the *same* as previously defined in (2.6) and (2.5), respectively, and the k^{th} longitudinal density is defined by either the Gaussian (2.1) or other candidate members of the exponential family (4.6)–(4.11).

The M-step at iteration $(m+1)$ is equivalently formed by maximising the sum of n sets of conditional expectations $\mathbb{E}\Big[g(\boldsymbol{b}_i)\,|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega}^{(m)}\Big]$, which are evaluated against the conditional distribution $f\big(\boldsymbol{b}_i | T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega}^{(m)}\big)$.

4.4.1 Parameter estimation via the approximate EM algorithm

As was outlined in Sections 4.2.2–4.2.3, we proceed with the approximation (3.1) since this allows us to steer clear of the (potentially) problematic density $f(\boldsymbol{b}_i | \boldsymbol{Y}_i; \boldsymbol{\Omega})$, which has intractable form outside of the multivariate normal case (2.18).

The approximate EM algorithm follows the same steps taken in Section 3.2.1, save for the starting values by `glmmTMB` (Brooks et al., 2017) now returning the starting values for dispersion parameter $\boldsymbol{\sigma}^{(0)} = \Big(\boldsymbol{\sigma}_1^{(0)^\top}, \ldots, \boldsymbol{\sigma}_K^{(0)^\top}\Big)^\top$, as the user can supply the '`dispformula`' argument here to specify the dispersion model (4.5) for each of the $k = 1, \ldots, K$ sub-

models.

Now, we consider the update to the parameter vector from iteration (m) to iteration $(m+1)$. We note that said update contains elements whose contribution to (4.12) (thereby their subsequent parameter update) does *not* change as we consider the more flexible case. Namely, the covariance matrix of the random effects D (whose update we outlined in Section 3.3.1); the baseline hazard $\lambda_0(\cdot)$ (Section 3.3.4); and survival parameters $\boldsymbol{\Phi}$ (Section 3.3.5). Therefore, we next consider the parameter updates for the fixed effects $\boldsymbol{\beta}^{(m)} \rightarrow \boldsymbol{\beta}^{(m+1)}$, and newly-introduced dispersion parameters $\boldsymbol{\sigma}^{(m)} \rightarrow \boldsymbol{\sigma}^{(m+1)}$.

4.4.2 The M-step for fixed effects $\boldsymbol{\beta}$

We first consider the parameter update for the k^{th} response's fixed effects $\boldsymbol{\beta}_k$ for each exponential family we outlined in Section 4.3.3. Irregardless of family chosen, we undertake a one-step Newton-Raphson iteration to update the parameter vector $\boldsymbol{\beta}_k^{(m)} \rightarrow \boldsymbol{\beta}_k^{(m+1)}$

$$\boldsymbol{\beta}_k^{(m+1)} = \boldsymbol{\beta}_k^{(m)} - \left[\sum_{i=1}^{n} \mathrm{H}_i\left(\boldsymbol{\beta}_k^{(m)}\right)\right]^{-1} \left[\sum_{i=1}^{n} s_i\left(\boldsymbol{\beta}_k^{(m)}\right)\right], \tag{4.13}$$

where $s_i\left(\boldsymbol{\beta}_k^{(m)}\right)$ is the gradient vector of the conditional expectation taken on the k^{th} complete data log-likelihood evaluated at the current estimate for the sub-model's fixed effects, and $\mathrm{H}_i\left(\boldsymbol{\beta}_k^{(m)}\right)$ is the Hessian matrix of second derivatives for subject i. Hereafter, we drop the subscript k for notational convenience.

In each of the proceeding sections – where we consider *family specific* parameter updates – we derive two quantities

$$\dot{\boldsymbol{\eta}}_i = \frac{\tilde{\mathbb{E}}_i[\ell_i(\boldsymbol{\eta}_i)]}{\partial \boldsymbol{\eta}_i}, \qquad \ddot{\boldsymbol{\eta}}_i = \frac{\tilde{\mathbb{E}}_i[\ell_i(\boldsymbol{\eta}_i)]}{\partial \boldsymbol{\eta}_i \partial \boldsymbol{\eta}_i}. \tag{4.14}$$

These two quantities allow us to form the score and Hessian in (4.13),

$$\begin{aligned} s_i\left(\boldsymbol{\beta}^{(m)}\right) &= \mathrm{X}_i^\top \dot{\boldsymbol{\eta}}_i, \\ \mathrm{H}_i\left(\boldsymbol{\beta}^{(m)}\right) &= \mathrm{X}_i^\top \operatorname{diag}(\ddot{\boldsymbol{\eta}}_i)\, \mathrm{X}_i, \end{aligned} \tag{4.15}$$

where $\operatorname{diag}(\ddot{\boldsymbol{\eta}}_i)$ is an $m_i \times m_i$ matrix with the values of the second derivative on its diagonal. The reason for undertaking this approach is it allows the author to escape more complex vector calculus, whilst achieving the same end result, by simply differentiating the quantity $\tilde{\mathbb{E}}_i[\ell_i(\boldsymbol{\eta}_i)]$ twice with respect to the linear predictor (essentially ignoring it being a vector).

For each family, we make use of the approximation (3.8), which we re-write here for

convenience's sake:

$$\boldsymbol{\eta}_i = \mathrm{X}_i\boldsymbol{\beta} + \mathrm{Z}_i\boldsymbol{b}_i \overset{\text{appx.}}{\sim} N\Big(\mathrm{X}_i\boldsymbol{\beta} + \mathrm{Z}_i\hat{\boldsymbol{b}}_i, \mathrm{Z}_i\hat{\Sigma}_i\mathrm{Z}_i^\top\Big) = N(\hat{\boldsymbol{\mu}}_i, \mathrm{A}_i)\,. \tag{4.16}$$

Poisson

The log-likelihood (4.6) w.r.t $\boldsymbol{\beta}$ (housed in linear predictor $\boldsymbol{\eta}_i$) is

$$\ell_i(\boldsymbol{\eta}_i) \underset{\boldsymbol{\beta}}{\propto} \boldsymbol{Y}_i^\top\boldsymbol{\eta}_i - \mathbf{1}^\top\exp\{\boldsymbol{\eta}_i\}$$

with expectation

$$\begin{aligned}\mathbb{E}_i[\ell_i(\boldsymbol{\eta}_i)] &= \mathbb{E}_i\Big[\boldsymbol{Y}_i^\top\boldsymbol{\eta}_i - \mathbf{1}^\top\exp\{\boldsymbol{\eta}_i\}\Big] \\ &= \boldsymbol{Y}_i^\top\mathbb{E}_i[\boldsymbol{\eta}_i] - \mathbf{1}^\top\mathbb{E}_i[\exp\{\boldsymbol{\eta}_i\}]\,.\end{aligned}$$

By (4.16) we obtain the approximate expectation

$$\tilde{\mathbb{E}}_i[\ell_i(\boldsymbol{\eta}_i)] = \boldsymbol{Y}_i^\top\hat{\boldsymbol{\mu}}_i - \mathbf{1}^\top\exp\big\{\hat{\boldsymbol{\mu}}_i + \boldsymbol{\tau}_i^2/2\big\}\,,$$

since the term $\exp\{\hat{\boldsymbol{\mu}}_i\}$ is log-normally distributed by (4.16). However, this log-normal term tends to be askew, such that appraisal at the mean $\exp\big\{\hat{\boldsymbol{\mu}}_i + \boldsymbol{\tau}_i^2/2\big\}$ is inappropriate; to which end we appraise instead using the median $\exp\{\hat{\boldsymbol{\mu}}_i\}$, which tends to be more centered in the peak of the resulting posterior distribution

$$\tilde{\mathbb{E}}_i[\ell_i(\boldsymbol{\eta}_i)] = \boldsymbol{Y}_i^\top\hat{\boldsymbol{\mu}}_i - \mathbf{1}^\top\exp\{\hat{\boldsymbol{\mu}}_i\}\,,$$

with justification given in Section 5.4. We can then obtain our two quantities of interest

$$\dot{\boldsymbol{\eta}}_i = \boldsymbol{Y}_i - \exp\{\hat{\boldsymbol{\mu}}_i\}\,, \qquad \ddot{\boldsymbol{\eta}}_i = -\exp\{\hat{\boldsymbol{\mu}}_i\}\,. \tag{4.17}$$

Binomial

The log-likelihood (4.7) w.r.t $\boldsymbol{\beta}$ is

$$\ell_i(\boldsymbol{\eta}_i) \underset{\boldsymbol{\beta}}{\propto} \boldsymbol{Y}_i^\top\boldsymbol{\eta}_i - \mathbf{1}^\top\log(\mathbf{1} + \exp\{\boldsymbol{\eta}_i\})\,,$$

with expectation

$$\begin{aligned}\mathbb{E}_i[\ell_i(\boldsymbol{\eta}_i)] &= \mathbb{E}_i\Big[\boldsymbol{Y}_i^\top\boldsymbol{\eta}_i - \mathbf{1}^\top\log(\mathbf{1} + \exp\{\boldsymbol{\eta}_i\})\Big] \\ &= \boldsymbol{Y}_i^\top\mathbb{E}_i[\boldsymbol{\eta}_i] - \mathbf{1}^\top\mathbb{E}_i[\log(\mathbf{1} + \exp\{\boldsymbol{\eta}_i\})]\,.\end{aligned}$$

Applying (4.16) we obtain the approximate expectation using Gauss-Hermite quadrature with ϱ weights and abscissae, as previously introduced in Section 3.3.2

$$\tilde{\mathbb{E}}_i[\ell_i(\boldsymbol{\eta}_i)] = \boldsymbol{Y}_i^\top \hat{\boldsymbol{\mu}}_i - \mathbf{1}^\top \sum_{l=1}^{\varrho} w_l \log(\mathbf{1} + \exp\{\hat{\boldsymbol{\mu}}_i + \boldsymbol{\tau}_i v_l\}) \,.$$

Finally, we obtain the two quantities of interest:

$$\begin{aligned} \dot{\boldsymbol{\eta}}_i &= \boldsymbol{Y}_i - \sum_{l=1}^{\varrho} w_l \frac{\exp\{\hat{\boldsymbol{\mu}}_i + \boldsymbol{\tau}_i v_l\}}{\mathbf{1} + \exp\{\hat{\boldsymbol{\mu}}_i + \boldsymbol{\tau}_i v_l\}}, \\ \ddot{\boldsymbol{\eta}}_i &= -\sum_{l=1}^{\varrho} w_l \frac{\exp\{\hat{\boldsymbol{\mu}}_i + \boldsymbol{\tau}_i v_l\}}{(\mathbf{1} + \exp\{\hat{\boldsymbol{\mu}}_i + \boldsymbol{\tau}_i v_l\})^2}. \end{aligned} \tag{4.18}$$

Negative binomial

The log-likelihood (4.8) pertaining to $\boldsymbol{\beta}$ is given by

$$\ell_i(\boldsymbol{\eta}_i) \underset{\boldsymbol{\beta}}{\propto} \boldsymbol{\varphi}_i^\top \log(\exp\{\boldsymbol{\eta}_i\} + \boldsymbol{\varphi}_i) + \boldsymbol{Y}_i^\top [\log \exp\{\boldsymbol{\eta}_i\} - \log(\exp\{\boldsymbol{\eta}_i\} + \boldsymbol{\varphi}_i)] \,,$$

with expectation

$$\begin{aligned} \mathbb{E}_i[\ell_i(\boldsymbol{\eta}_i)] &= \mathbb{E}_i\Big[\boldsymbol{\varphi}_i^\top \log(\exp\{\boldsymbol{\eta}_i\} + \boldsymbol{\varphi}_i) + \boldsymbol{Y}_i^\top \boldsymbol{\eta}_i - \boldsymbol{Y}_i^\top \log(\exp\{\boldsymbol{\eta}_i\} + \boldsymbol{\varphi}_i)\Big] \\ &= \boldsymbol{\varphi}_i^\top \mathbb{E}_i[\log(\exp\{\boldsymbol{\eta}_i\} + \boldsymbol{\varphi}_i)] + \boldsymbol{Y}_i^\top \mathbb{E}_i[\boldsymbol{\eta}_i] - \boldsymbol{Y}_i^\top \mathbb{E}_i[\log(\exp\{\boldsymbol{\eta}_i\} + \boldsymbol{\varphi}_i)] \\ &= \boldsymbol{Y}_i^\top \mathbb{E}_i[\boldsymbol{\eta}_i] + (\boldsymbol{\varphi}_i - \boldsymbol{Y}_i)^\top \mathbb{E}_i[\log(\exp\{\boldsymbol{\eta}_i\} + \boldsymbol{\varphi}_i)] \,. \end{aligned}$$

Next, appraising via (4.16) we obtain the approximate expectation

$$\tilde{\mathbb{E}}_i[\ell_i(\boldsymbol{\eta}_i)] = \boldsymbol{Y}_i^\top \hat{\boldsymbol{\mu}}_i + (\boldsymbol{\varphi}_i - \boldsymbol{Y}_i)^\top \sum_{l=1}^{\varrho} w_l \log(\exp\{\hat{\boldsymbol{\mu}}_i + \boldsymbol{\tau}_i v_l\} + \boldsymbol{\varphi}_i) \,,$$

before finally the first and second derivatives with respect to the linear predictor $\boldsymbol{\eta}_i$,

$$\begin{aligned} \dot{\boldsymbol{\eta}}_i &= \boldsymbol{Y}_i - (\boldsymbol{Y}_i + \boldsymbol{\varphi}_i) \sum_{l=1}^{\varrho} w_l \frac{\exp\{\hat{\boldsymbol{\mu}}_i + \boldsymbol{\tau}_i v_l\}}{\boldsymbol{\varphi}_i + \exp\{\hat{\boldsymbol{\mu}}_i + \boldsymbol{\tau}_i v_l\}}, \\ \ddot{\boldsymbol{\eta}}_i &= -(\boldsymbol{Y}_i + \boldsymbol{\varphi}_i) \sum_{l=1}^{\varrho} w_l \frac{\boldsymbol{\varphi}_i \exp\{\hat{\boldsymbol{\mu}}_i + \boldsymbol{\tau}_i v_l\}}{(\boldsymbol{\varphi}_i + \exp\{\hat{\boldsymbol{\mu}}_i + \boldsymbol{\tau}_i v_l\})^2}. \end{aligned} \tag{4.19}$$

Generalised Poisson

The log-likelihood (4.9) pertaining to $\boldsymbol{\beta}$ is given by

$$\ell_i(\boldsymbol{\eta}_i) \underset{\boldsymbol{\beta}}{\propto} \mathbf{1}^\top \boldsymbol{\eta}_i + (\boldsymbol{Y}_i - \mathbf{1})^\top \log(\exp\{\boldsymbol{\eta}_i\} + \boldsymbol{\varphi}_i \boldsymbol{Y}_i) - \mathbf{1}^\top \frac{\exp\{\boldsymbol{\eta}_i\}}{\mathbf{1} + \boldsymbol{\varphi}_i},$$

with expectation

$$\begin{aligned}\mathbb{E}_i[\ell_i(\boldsymbol{\eta}_i)] &= \mathbb{E}_i\left[\mathbf{1}^\top \boldsymbol{\eta}_i + (\boldsymbol{Y}_i - \mathbf{1})^\top \log(\exp\{\boldsymbol{\eta}_i\} + \boldsymbol{\varphi}_i \boldsymbol{Y}_i) - \mathbf{1}^\top \frac{\exp\{\boldsymbol{\eta}_i\}}{\mathbf{1} + \boldsymbol{\varphi}_i}\right] \\ &= \mathbf{1}^\top \mathbb{E}_i[\boldsymbol{\eta}_i] + (\boldsymbol{Y}_i - \mathbf{1})^\top \mathbb{E}_i[\log(\exp\{\boldsymbol{\eta}_i\} + \boldsymbol{\varphi}_i \boldsymbol{Y}_i)] - \mathbf{1}^\top \frac{\mathbb{E}_i[\exp\{\boldsymbol{\eta}_i\}]}{\mathbf{1} + \boldsymbol{\varphi}_i},\end{aligned}$$

which we evaluate by approximation (4.16)

$$\tilde{\mathbb{E}}_i[\ell_i(\boldsymbol{\eta}_i)] = \mathbf{1}^\top \hat{\boldsymbol{\mu}}_i - \mathbf{1}^\top \frac{\exp\{\hat{\boldsymbol{\mu}}_i\}}{\mathbf{1} + \boldsymbol{\varphi}_i} + (\boldsymbol{Y}_i - \mathbf{1})^\top \sum_{l=1}^{\varrho} w_l \log(\exp\{\hat{\boldsymbol{\mu}}_i + \boldsymbol{\tau}_i v_l\} + \boldsymbol{\varphi}_i \boldsymbol{Y}_i)\,.$$

In this approximate expectation, we once more appraise the expected value of the log-normal quantity, $\mathbb{E}_i[\exp\{\hat{\boldsymbol{\mu}}_i\}]$, by the median of the log-normal distribution to escape issues surrounding the mean of the arising log-normal distribution usually being located away from its modal mass, with some justification provided in Section 5.4. Finally, the two quantities necessary for updating $\boldsymbol{\beta}$ are

$$\begin{aligned}\dot{\boldsymbol{\eta}}_i &= \mathbf{1} - \frac{\exp\{\hat{\boldsymbol{\mu}}_i\}}{\mathbf{1} + \boldsymbol{\varphi}_i} + (\boldsymbol{Y}_i - \mathbf{1}) \sum_{l=1}^{\varrho} w_l \frac{\exp\{\hat{\boldsymbol{\mu}}_i + \boldsymbol{\tau}_i v_l\}}{\exp\{\hat{\boldsymbol{\mu}}_i + \boldsymbol{\tau}_i v_l\} + \boldsymbol{\varphi}_i \boldsymbol{Y}_i}, \\ \ddot{\boldsymbol{\eta}}_i &= -\frac{\exp\{\hat{\boldsymbol{\mu}}_i\}}{\mathbf{1} + \boldsymbol{\varphi}_i} + (\boldsymbol{Y}_i - \mathbf{1}) \sum_{l=1}^{\varrho} w_l \frac{\boldsymbol{\varphi}_i \boldsymbol{Y}_i \exp\{\hat{\boldsymbol{\mu}}_i + \boldsymbol{\tau}_i v_l\}}{(\exp\{\hat{\boldsymbol{\mu}}_i + \boldsymbol{\tau}_i v_l\} + \boldsymbol{\varphi}_i \boldsymbol{Y}_i)^2}\end{aligned} \tag{4.20}$$

Gamma

The log-likelihood (4.11) which involves $\boldsymbol{\beta}$ is

$$\begin{aligned}\ell_i(\boldsymbol{\eta}_i) &\underset{\boldsymbol{\beta}}{\propto} -\mathbf{1}^\top \left(\frac{\boldsymbol{Y}_i \boldsymbol{\varphi}_i}{\exp\{\boldsymbol{\eta}_i\}}\right) - \boldsymbol{\varphi}_i^\top \log\left(\frac{\exp\{\boldsymbol{\eta}_i\}}{\boldsymbol{\varphi}_i}\right), \\ &= -\mathbf{1}^\top \left(\frac{\boldsymbol{Y}_i \boldsymbol{\varphi}_i}{\exp\{\boldsymbol{\eta}_i\}}\right) - \boldsymbol{\varphi}_i^\top \boldsymbol{\eta}_i,\end{aligned}$$

with expectation

$$\begin{aligned}\mathbb{E}_i[\ell_i(\boldsymbol{\eta}_i)] &= \mathbb{E}_i\left[-\mathbf{1}^\top\left(\frac{\boldsymbol{Y}_i\boldsymbol{\varphi}_i}{\exp\{\boldsymbol{\eta}_i\}}\right) - \boldsymbol{\varphi}_i^\top\boldsymbol{\eta}_i\right] \\ &= -\mathbf{1}^\top\left(\frac{\boldsymbol{Y}_i\boldsymbol{\varphi}_i}{\mathbb{E}_i[\exp\{\boldsymbol{\eta}_i\}]}\right) - \boldsymbol{\varphi}_i^\top\mathbb{E}_i[\boldsymbol{\eta}_i]\,,\end{aligned}$$

which we can once more evaluate by using approximation (4.16)

$$\tilde{\mathbb{E}}_i[\ell_i(\boldsymbol{\eta}_i)] = -\mathbf{1}^\top\left(\frac{\boldsymbol{Y}_i\boldsymbol{\varphi}_i}{\exp\{\hat{\boldsymbol{\mu}}_i\}}\right) - \boldsymbol{\varphi}_i^\top\hat{\boldsymbol{\mu}}_i.$$

Fairly trivially then we obtain our two key quantities:

$$\dot{\boldsymbol{\eta}}_i = \boldsymbol{\varphi}_i(\exp\{-\hat{\boldsymbol{\mu}}_i\}\,\boldsymbol{Y}_i - \mathbf{1})\,, \qquad \ddot{\boldsymbol{\eta}}_i = -\exp\{-\hat{\boldsymbol{\mu}}_i\}\,\boldsymbol{Y}_i\boldsymbol{\varphi}_i. \tag{4.21}$$

4.4.3 The M-step for dispersion parameters $\boldsymbol{\sigma}$

We now consider updates to the dispersion parameters where applicable. We note that the residual variance $\sigma^2_{\varepsilon_k} \in \boldsymbol{\sigma}$ for the k^{th} Gaussian response remains unchanged from Section 3.3.3. In proceeding sections, we only present derivations of the quantity $\tilde{\mathbb{E}}_i[\ell_i(\boldsymbol{\sigma}_k)]$, and note that we obtain $s_i(\boldsymbol{\sigma}_k)$ and $\mathrm{H}_i(\boldsymbol{\sigma}_k)$ $\forall\ i = 1, \ldots, n$ using central differencing, which we outline in Appendix A.1. We can use these quantities to form the Newton-Raphson iteration to update $\boldsymbol{\sigma}_k^{(m)} \rightarrow \boldsymbol{\sigma}_k^{(m+1)}$ in an manner analogous to (4.13), namely

$$\boldsymbol{\sigma}_k^{(m+1)} = \boldsymbol{\sigma}_k^{(m)} - \left[\sum_{i=1}^{n}\mathrm{H}_i\Big(\boldsymbol{\sigma}_k^{(m)}\Big)\right]^{-1}\left[\sum_{i=1}^{n}s_i\Big(\boldsymbol{\sigma}_k^{(m)}\Big)\right], \tag{4.22}$$

if the k^{th} family is either Negative binomial, Generalised Poisson or Gamma. In the next sub-sections, we present the approximate expectations (sans k-notation) upon which we carry out the numerical differentiation routines. The dispersion parameter(s) $\boldsymbol{\sigma}$ is included in requisite expressions by $\boldsymbol{\varphi}_i$ through some known link function $\tilde{h}(\cdot)$ given in Table 4.1.

Negative binomial

The contribution of dispersion parameter $\boldsymbol{\sigma}$ to the negative binomial log-likelihood (4.8) is given by

$$\ell_i(\boldsymbol{\sigma}) \underset{\boldsymbol{\sigma}}{\propto} \mathbf{1}^\top\{\log\Gamma(\boldsymbol{Y}_i + \boldsymbol{\varphi}_i) - \log\Gamma(\boldsymbol{\varphi}_i)\} + \boldsymbol{\varphi}_i^\top\log\boldsymbol{\varphi}_i - (\boldsymbol{\varphi}_i + \boldsymbol{Y}_i)^\top\log(\exp\{\boldsymbol{\eta}_i\} + \boldsymbol{\varphi}_i)\,,$$

which has expectation

$$\mathbb{E}_i[\ell_i(\boldsymbol{\sigma})] = \mathbf{1}^\top \{\log\Gamma(\boldsymbol{Y}_i + \boldsymbol{\varphi}_i) - \log\Gamma(\boldsymbol{\varphi}_i)\} + \boldsymbol{\varphi}_i^\top \log\boldsymbol{\varphi}_i \\ -(\boldsymbol{\varphi}_i + \boldsymbol{Y}_i)^\top \mathbb{E}_i[\log(\exp\{\boldsymbol{\eta}_i\} + \boldsymbol{\varphi}_i)]\,,$$

which can be appraised using the approximation (4.16), resulting in the required expectation

$$\tilde{\mathbb{E}}_i[\ell_i(\boldsymbol{\sigma})] = \mathbf{1}^\top \{\log\Gamma(\boldsymbol{Y}_i + \boldsymbol{\varphi}_i) - \log\Gamma(\boldsymbol{\varphi}_i)\} + \boldsymbol{\varphi}_i^\top \log\boldsymbol{\varphi}_i \\ -(\boldsymbol{\varphi}_i + \boldsymbol{Y}_i)^\top \sum_{l=1}^{\varrho} w_l \log(\exp\{\hat{\boldsymbol{\mu}}_i + \boldsymbol{\tau}_i v_l\} + \boldsymbol{\varphi}_i)\,. \tag{4.23}$$

Generalised Poisson

The corresponding portion of the log-likelihood (4.9) is

$$\ell_i(\boldsymbol{\sigma}) \underset{\boldsymbol{\sigma}}{\propto} (\boldsymbol{Y}_i - \mathbf{1})^\top \log(\exp\{\boldsymbol{\eta}_i\} + \boldsymbol{\varphi}_i \boldsymbol{Y}_i) - \boldsymbol{Y}_i^\top \log(\mathbf{1} + \boldsymbol{\varphi}_i) - \mathbf{1}^\top \left(\frac{\exp\{\boldsymbol{\eta}_i\} + \boldsymbol{\varphi}_i \boldsymbol{Y}_i}{\mathbf{1} + \boldsymbol{\varphi}_i}\right),$$

with expectation

$$\mathbb{E}_i[\ell_i(\boldsymbol{\sigma})] = (\boldsymbol{Y}_i - \mathbf{1})^\top \mathbb{E}_i[\log(\exp\{\boldsymbol{\eta}_i\} + \boldsymbol{\varphi}_i \boldsymbol{Y}_i)] - \boldsymbol{Y}_i^\top \log(\mathbf{1} + \boldsymbol{\varphi}_i) \\ -\left(\frac{\mathbf{1}}{\mathbf{1} + \boldsymbol{\varphi}_i}\right)^\top \big(\mathbb{E}_i[\exp\{\boldsymbol{\eta}_i\}] + \boldsymbol{\varphi}_i \boldsymbol{Y}_i\big).$$

Finally, applying the approximation on the linear predictor (4.16) to obtain the approximated expectation

$$\tilde{\mathbb{E}}_i[\ell_i(\boldsymbol{\sigma})] = (\boldsymbol{Y}_i - \mathbf{1})^\top \sum_{l=1}^{\varrho} w_l \log(\exp\{\hat{\boldsymbol{\mu}}_i + \boldsymbol{\tau}_i v_l\} + \boldsymbol{\varphi}_i \boldsymbol{Y}_i) - \boldsymbol{Y}_i^\top \log(\mathbf{1} + \boldsymbol{\varphi}_i) \\ -\left(\frac{\mathbf{1}}{\mathbf{1} + \boldsymbol{\varphi}_i}\right)^\top \big(\exp\{\hat{\boldsymbol{\mu}}_i\} + \boldsymbol{\varphi}_i \boldsymbol{Y}_i\big). \tag{4.24}$$

Gamma

The requisite part of the log-likelihood (4.11) is

$$\ell_i(\boldsymbol{\sigma}) \underset{\boldsymbol{\sigma}}{\propto} (\boldsymbol{\varphi}_i - \mathbf{1})^\top \boldsymbol{Y}_i - \mathbf{1}^\top \left(\frac{\boldsymbol{Y}_i \boldsymbol{\varphi}_i}{\exp\{\boldsymbol{\eta}_i\}}\right) - \mathbf{1}^\top \log\Gamma(\boldsymbol{\varphi}_i) + \boldsymbol{\varphi}_i^\top \log\boldsymbol{\varphi}_i,$$

with expectation

$$\mathbb{E}_i[\ell_i(\boldsymbol{\sigma})] = (\boldsymbol{\varphi}_i - \mathbf{1})^\top \boldsymbol{Y}_i - \mathbf{1}^\top \left(\frac{\boldsymbol{Y}_i \boldsymbol{\varphi}_i}{\mathbb{E}_i[\exp\{\boldsymbol{\eta}_i\}]} \right) - \mathbf{1}^\top \log \Gamma(\boldsymbol{\varphi}_i) + \boldsymbol{\varphi}_i^\top \log \boldsymbol{\varphi}_i,$$

which we appraise using (4.16)

$$\tilde{\mathbb{E}}_i[\ell_i(\boldsymbol{\sigma})] = (\boldsymbol{\varphi}_i - \mathbf{1})^\top \boldsymbol{Y}_i - \mathbf{1}^\top \left(\frac{\boldsymbol{Y}_i \boldsymbol{\varphi}_i}{\exp\{\hat{\boldsymbol{\mu}}_i\}} \right) - \mathbf{1}^\top \log \Gamma(\boldsymbol{\varphi}_i) + \boldsymbol{\varphi}_i^\top \log \boldsymbol{\varphi}_i. \tag{4.25}$$

4.5 Simulation studies

With performance of the approximate EM algorithm established under multiple scenarios in Section 3.4, we don't consider quite as many situations in this chapter, assuming performance of the algorithm will remain largely unchanged under different data types. Instead, we undertake simulations which illustrate the parameter estimation capabilities for these non-Gaussian response distributions. To that end, we consider trivariate and five-variate scenarios consisting of the three 'de facto' distributions for continuous, count and binary data, followed by a series of *univariate* sets of joint models fit to simulated data from the more 'non-standard' distributions. The aim of these univariate simulations is to show good estimation capabilities afforded by the approximate EM algorithm.

In a similar vein to how the 'default' simulation scenario was laid out in the purely Gaussian case in Section 3.4.1, we outline those default parameter values in the next section; followed by a brief note on the actual simulation procedures used for each of these families.

4.5.1 Default values for simulation of a flexible longitudinal processes

Here, we outline the choices for $\boldsymbol{\Omega}^{(\text{TRUE})}_{\mathrm{D}_k,\boldsymbol{\beta}_k,\boldsymbol{\sigma}_k}$. Departing from setting of the true parameter based on the longitudinal response 'number' $k = 1, \ldots, K$, we now set the true parameter values for the k^{th} response depending on the family. The true parameter values for the longitudinal sub-model for $h_k\Big(\mathbb{E}_i\Big[\boldsymbol{Y}_{ik}\Big|\boldsymbol{b}_{ik}; \boldsymbol{\Omega}^{(\text{TRUE})}_{\mathrm{D}_k,\boldsymbol{\beta}_k,\boldsymbol{\sigma}_k}\Big]\Big)$ are shown in Table 4.2. The full covariance matrix (i.e. for the multivariate joint model) is defined in the same way $\mathrm{D} = \bigoplus_{k=1}^{K} \mathrm{D}_k$ with covariance between each random intercept set as 0.125.

The chosen parameter values for Poisson and negative binomial count responses are chosen to avoid situations where the simulated response has an excess of zeroes. The parameter choices for the Gamma case ensures a right skewed distribution which precludes exceedingly large values in its long tail. We consider two-fold data generation of *both* under- and

Response distribution	$\boldsymbol{\beta}_k^\top$	$\text{diag}(\text{D}_k)^\top$	$\boldsymbol{\sigma}_k^\top$
Gaussian	$(-2.00, 0.10, -0.10, 0.20)$	$(0.25, 0.09)$	$\equiv \sigma^2_{\varepsilon_k} = 0.16$
Gamma	$(0.00, -0.10, 0.10, -0.20)$	$(0.20, 0.05)$	2.00
Poisson	$(2.00, -0.10, 0.10, 0.20)$	$(0.50, 0.09)$	N/A
Negative binomial	$(2.00, -0.10, 0.10, 0.20)$	$(0.50, 0.09)$	1.00
Generalised Poisson (under)	$(1.00, 0.05, -0.05, 0.10)$	$(0.30, 0.00)$	-0.30
Generalised Poisson (over)	$(0.50, -0.20, 0.05, 0.40)$	$(0.70, 0.00)$	0.30
Binomial	$(1.00, -1.00, 1.00, -1.00)$	$(2.00, 0.00)$	N/A

Table 4.2: Choices for $\boldsymbol{\Omega}^{(\text{TRUE})}_{\text{D}_k,\boldsymbol{\beta}_k,\boldsymbol{\sigma}_k}$ for a chosen distribution on the k^{th} longitudinal response. *N/A*: Not applicable i.e. no dispersion to be modelled. For the generalised Poisson, 'under': Underdispersed; 'over': Overdispersed.

overdispersed data under the generalised Poisson.

The simulation of survival times follows the same process as previously outlined in Section 3.4.1; we re-emphasise that whilst elements of $\boldsymbol{\beta}$, D and $\boldsymbol{\sigma}$ now depend on the *distribution*, the corresponding k^{th} element of $\boldsymbol{\gamma}$ still depends *only* on whether k is odd or even. The baseline hazard is controlled to ensure approximately 30% failure rate across all $N = 300$ simulated datasets.

Remark. Notably, in the random effect specification under a binary response a random intercept is chosen over the slope used for all other response types. This was because (perhaps due to truncation of longitudinal profiles) it was challenging to reliably generate binomial data which `glmmTMB` was able to fit to in the intended way. The same reasoning extends to the generalised Poisson.

4.5.2 Note on simulation from candidate response distributions

In order to 'complete the picture' in anticipation of subsequent simulations to be carried out, we briefly expand upon the routine outlined in Section 2.5.1 for generation of (Gaussian) longitudinal responses.

We define the vector of true parameter values $\boldsymbol{\Omega}^{(\text{TRUE})}_{\text{D},\boldsymbol{\beta},\boldsymbol{\sigma}} = \left(\text{vech}(\text{D})^\top, \boldsymbol{\beta}^\top, \boldsymbol{\sigma}^\top\right)^\top$ with which we form the linear predictor for the longitudinal process, $\boldsymbol{\eta}_i = \text{X}_i\boldsymbol{\beta} + \text{Z}_i\boldsymbol{b}_i$, and dispersion process $\tilde{\boldsymbol{\eta}}_i = \text{W}_i\boldsymbol{\sigma}$ (see Section 4.3.2 for details and special cases), with W_i an appropriately-generated design matrix. Simulation of the random effects $\boldsymbol{b}_i$ is unchanged, as is the simulation of survival times.

With linear predictor $\boldsymbol{\eta}_i$ in-hand, we turn attention to simulation of the longitudinal response through inverse link function: $\mathbb{E}_i\left[\boldsymbol{Y}_i|\boldsymbol{b}_i; \boldsymbol{\Omega}^{(\text{TRUE})}_{\text{D},\boldsymbol{\beta},\boldsymbol{\sigma}}\right] = h^{-1}(\boldsymbol{\eta}_i)$ where $h(\cdot)$ is defined for each candidate distribution considered in Table 4.1. In practice, built-in `R` functions are used for the Poisson `rpois`; binomial `rbinom` with argument `size` set to 1; Gamma `rgamma` with arguments `shape` set as $\exp\{\tilde{\boldsymbol{\eta}}_i\}$ and `scale` as $\exp\{\boldsymbol{\eta}_i\} \,/\, \exp\{\tilde{\boldsymbol{\eta}}_i\}$; negative

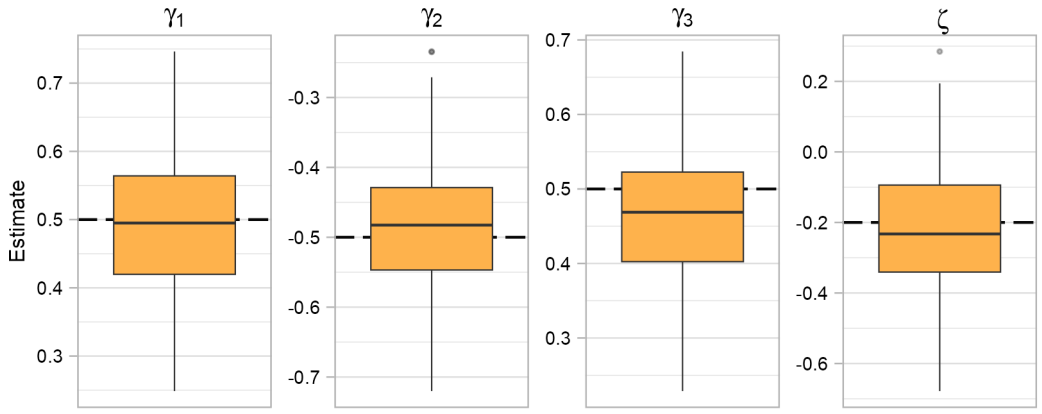


Figure 4.1: Estimates for survival parameters $\hat{\boldsymbol{\gamma}}$ and $\hat{\zeta}$ for a trivariate mixture joint model. The dashed line signifies the true parameter value.

binomial using `rnegbin` from the `MASS` package (Venables and Ripley, 2002); and finally generalised Poisson using a bespoke simulation function.

4.5.3 Trivariate mixture of families

Beginning with a trivariate joint model with one Gaussian ($\boldsymbol{Y}_{i1}$), one Poisson ($\boldsymbol{Y}_{i2}$) and one binary ($\boldsymbol{Y}_{i3}$) longitudinal sub-model, we demonstrate parameter estimation capabilities for a realistic scenario where three different types of data are jointly modelled using their 'most natural' or de facto distribution.

Figure 4.1 illustrates estimation for the survival parameters $\hat{\boldsymbol{\Phi}}$. We note estimates of the association parameters $\hat{\boldsymbol{\gamma}}$ are well-estimated, but there appears to be some overestimation in $\hat{\gamma}_2$ (associated with the Poisson) as well as underestimation in $\hat{\gamma}_3$ (binomial). The estimates for $\hat{\zeta}$ appears to have suffered slightly in light of these two association parameters.

The fully tabulated results are presented in Appendix B.2.1, where we note broadly good performance. However, the variance of the random intercept for the binomial response, $\hat{\mathrm{D}}_{5,5}$, and the fixed effect intercept $\hat{\beta}_{30}$ suffer poorest coverage. This perhaps indicating some conflation between the two; the model perhaps unable to attribute variance in the binomial response to the random effects.

4.5.4 Further trivariate simulations

Mimicking the investigations carried out into the effect of the follow-up length r and failure rate ω in Sections 3.4.4 and 3.4.5, respectively, we monitor these in a joint fashion for the trivariate mixture as considered in the previous section. Utilising the same candidate

values for $r \in \{5, 10, 15\}$, $\omega \in \{10\%, 30\%, 50\%\}$, we undertake $N = 200$ simulations for each combination of $\{r\} \times \{\omega\}$ with simulated sample size $n = 250$.

The parameter estimates for the survival parameters $\hat{\boldsymbol{\gamma}}$ and $\hat{\zeta}$ are given in Figure 4.2. Here, we observe once more that an increased number of failure times leads to a reduction in spread about the estimates for $\hat{\zeta}$, which is well-estimated with relatively little bias. The time-varying survival parameters $\hat{\boldsymbol{\gamma}}$ are well-estimated overall. Generally speaking, we note a reduction in bias as r increases from 5 time-points. However, we note routine underestimation in $\hat{\gamma}_3$, though we do not expect this to improve with longer profiles, since it is attached to the random intercept only.

Additional results are provided in Appendix B.2.2. We observe broadly good coverage from $\text{vech}(\hat{\mathrm{D}})$, apart from $\hat{\mathrm{D}}_{55}$ – attached to the binary sub-model – which routinely has poorest coverage. There does appear to be some systematic underestimation of these variance-covariance components; this bias decreasing as r increases. The fixed effects $\hat{\boldsymbol{\beta}}$ also appear to enjoy reduction in bias as the profile length increases, with the intercept attached to the binary sub-model suffering the poorest coverage across simulations, with all other components having coverage around the nominal 0.95. Finally we observe the average estimated standard error generally gets closer to the empirical standard deviation as r increases, this suggesting that a longer profile allows the approach to more accurately capture the variability surrounding these parameter estimates.

The elapsed time for convergence of the approximate EM algorithm and additional time taken for calculation of standard errors is given in Table 4.3, where we note a longer profile reduces this computation time, and a larger proportion of failures increases it; this phenomena was previously remarked upon in Section 3.4.5.

	$r = 5$	$r = 10$	$r = 15$
$\omega = 10\%$	4.628 [4.125, 5.237]	3.217 [2.896, 3.608]	3.175 [2.828, 3.631]
$\omega = 30\%$	5.868 [5.167, 6.916]	4.110 [3.720, 4.577]	3.964 [3.620, 4.383]
$\omega = 50\%$	7.422 [6.520, 8.624]	5.237 [4.771, 5.801]	4.819 [4.380, 5.412]

Table 4.3: Median [IQR] elapsed time in seconds for simulations considered in the trivariate simulations. The proportion who fail is denoted by ω and the maximal length of the longitudinal profiles by r.

4.5.5 Five-variate mixture joint model

With performative capabilities laid out in a trivariate setting – across differing simulation scenarios – in the previous section, we seek to offer an idea of performance of the algorithm for both a larger sample size $n = 500$ and a greater number of longitudinal responses $K = 5$ for a larger simulated sample $N = 500$. Specifically here we opt for two Gaussian ($k = 1, 2$), two Poisson ($k = 3, 4$), and one binomial ($k = 5$) longitudinal responses.

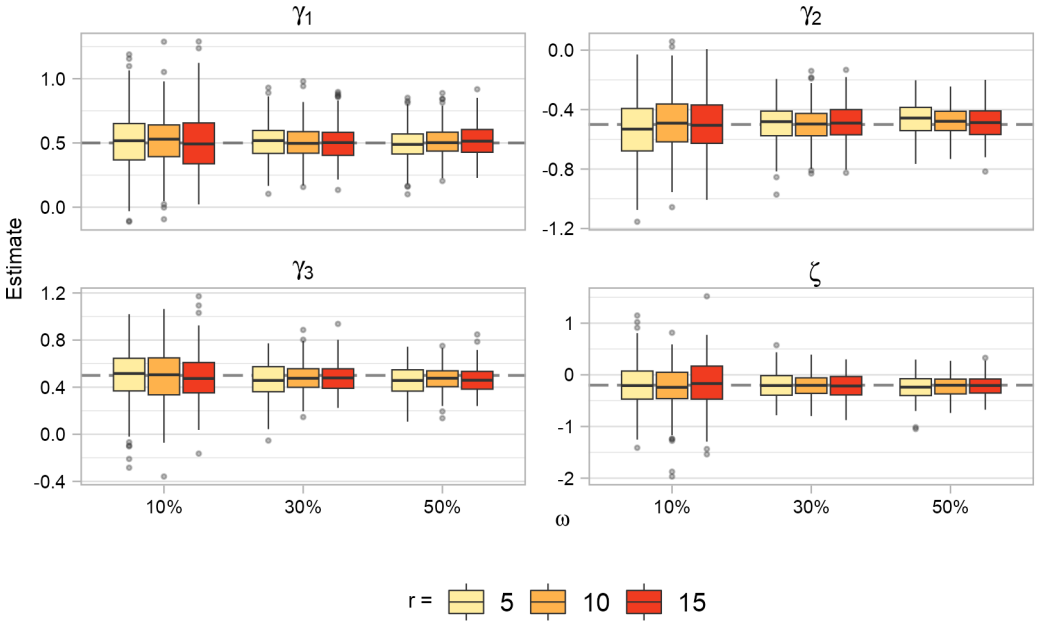


Figure 4.2: Estimates for survival parameters $\hat{\boldsymbol{\gamma}}$ and $\hat{\zeta}$ for differing maximal profiles lengths r and failure rates ω. The dashed line signifies the true parameter value.

We opt for slightly different true values to those given in 4.5.1, and set $\text{diag}(\text{D}) = (0.25, 0.09, 0.30, 0.06, 0.20, 0.04, 0.50, 0.09, 2.00)^\top$ with resulting correlation between random intercepts of the e^{th} and f^{th} responses ρ_{ef}: $\rho_{13} = 0.46$, $\rho_{15} = 0.56$, $\rho_{17} = 0.35$, $\rho_{19} = 0.18$, $\rho_{35} = 0.51$, $\rho_{37} = 0.32$, $\rho_{39} = 0.16$, $\rho_{57} = 0.40$, $\rho_{59} = 0.20$, $\rho_{79} = 0.13$. Additionally, we mute slightly the association parameters $\boldsymbol{\gamma} = (0.25, -0.25, 0.25, -0.25, 0.30)^\top$.

The estimates for survival parameters $\hat{\boldsymbol{\gamma}}$ and $\hat{\zeta}$ are presented in Figure 4.3 where we note good estimation of all five association parameters as well as the time-invariant $\hat{\zeta}$. We note underestimation amongst $\hat{\gamma}_5$, and overestimation in $\hat{\gamma}_3$ with percentage biases of (-)12.0% and 11.6%, respectively. The median [IQR] elapsed time for the approximate EM algorithm to converge and standard errors calculated was 20.684 [19.358, 22.491] seconds; the escalation in computing times that hampers traditional quadrature approaches is diluted under the proposed approach.

A full summary of all $N = 500$ sets of model fits are given in Table 4.4. Generally speaking, we do not observe deterioration in estimation capabilities when considering two additional longitudinal responses, though the estimates (and/or their estimated standard errors) could be deemed conservative. In general, the average estimated standard error is in-line with the empirical standard deviation of parameter estimates, indicating a broadly good handle on the variability of the parameters; notably on the whole these are slightly overestimated, opposing the underestimation one expects given earlier results in literature

(Hsieh et al., 2006). However, as noted in Appendix A.3, a larger sample, incidence rate, and/or other data characteristics may be necessary for adequately establishing a handle here.

As was the case with the trivariate scenario, we note the fixed effect intercept, as well as the variance of the random intercept for the binomial sub-model suffers the poorest coverage. The remaining variance-covariance components $\text{vech}\big(\hat{\text{D}}\big)$ appear to be well estimated, but we once more note systematic *underestimation* of these terms; something that may improve by alterations to convergence criterion (3.4).

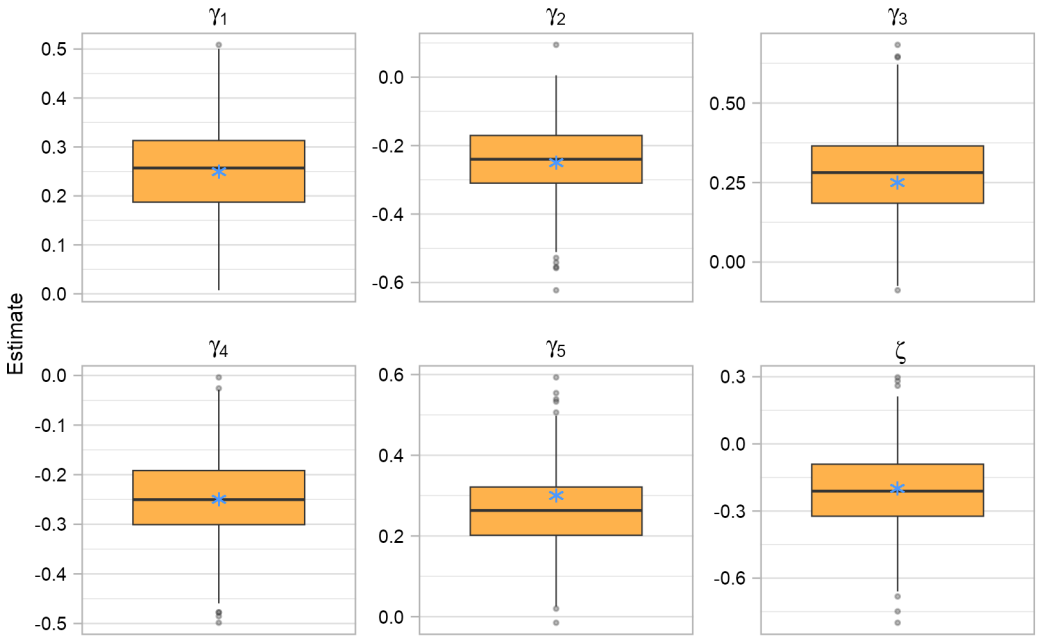


Figure 4.3: Estimates for survival parameters $\hat{\boldsymbol{\gamma}}$ and $\hat{\zeta}$ for the five-variate mixture model. A blue asterisk (*) signifies the true parameter value.

4.5.6 Univariate joint models: Gamma; negative binomial; and generalised Poisson

We now consider four univariate joint models with GLMM sub-model specified by the 'non-standard' distributions where dispersion is additionally modelled: One Gamma, one negative binomial and both under- and over-dispersed generalised Poisson.

Figure 4.4 presents the estimates for survival parameters $\hat{\boldsymbol{\Phi}}$ obtained for each longitudinal specification, where we note that the different longitudinal specifications don't appear to detract from the good performance noted in the purely Gaussian case. Here, the underdispersed generalised Poisson has notably the largest spread, which is likely attributable to

Parameter	Emp. Mean (SD)	Mean SE	Bias	MSE	CP
$D_{11} = 0.250$	0.248 (0.020)	0.022	-0.002	0.000	0.962
$D_{31} = 0.125$	0.125 (0.015)	0.018	0.000	0.000	0.974
$D_{51} = 0.125$	0.123 (0.014)	0.016	-0.002	0.000	0.960
$D_{71} = 0.125$	0.123 (0.019)	0.022	-0.002	0.000	0.976
$D_{91} = 0.125$	0.117 (0.043)	0.049	-0.008	0.002	0.970
$D_{22} = 0.090$	0.090 (0.007)	0.008	0.000	0.000	0.968
$D_{33} = 0.300$	0.299 (0.022)	0.026	-0.001	0.000	0.956
$D_{53} = 0.125$	0.125 (0.014)	0.017	0.000	0.000	0.970
$D_{73} = 0.125$	0.123 (0.021)	0.023	-0.002	0.000	0.964
$D_{93} = 0.125$	0.119 (0.047)	0.053	-0.006	0.002	0.966
$D_{44} = 0.060$	0.060 (0.005)	0.005	0.000	0.000	0.968
$D_{55} = 0.200$	0.198 (0.017)	0.019	-0.002	0.000	0.970
$D_{75} = 0.125$	0.122 (0.017)	0.020	-0.003	0.000	0.980
$D_{95} = 0.125$	0.117 (0.039)	0.045	-0.008	0.002	0.974
$D_{66} = 0.040$	0.039 (0.003)	0.004	-0.001	0.000	0.942
$D_{77} = 0.500$	0.484 (0.034)	0.041	-0.016	0.001	0.942
$D_{97} = 0.125$	0.119 (0.059)	0.065	-0.006	0.004	0.962
$D_{88} = 0.090$	0.087 (0.007)	0.008	-0.003	0.000	0.942
$D_{99} = 2.000$	1.696 (0.210)	0.267	-0.304	0.137	0.784
$\beta_{10} = 2.000$	1.990 (0.035)	0.039	-0.010	0.001	0.966
$\beta_{11} = -0.100$	-0.100 (0.015)	0.017	0.000	0.000	0.972
$\beta_{12} = 0.100$	0.100 (0.026)	0.028	0.000	0.001	0.960
$\beta_{13} = -0.200$	-0.200 (0.051)	0.054	0.000	0.003	0.964
$\beta_{20} = -2.000$	-2.007 (0.038)	0.042	-0.007	0.001	0.962
$\beta_{21} = 0.100$	0.103 (0.012)	0.014	0.003	0.000	0.976
$\beta_{22} = -0.100$	-0.098 (0.027)	0.030	0.002	0.001	0.960
$\beta_{23} = 0.200$	0.200 (0.053)	0.058	0.000	0.003	0.970
$\beta_{30} = 2.000$	1.998 (0.032)	0.035	-0.002	0.001	0.966
$\beta_{31} = -0.100$	-0.101 (0.011)	0.012	-0.001	0.000	0.974
$\beta_{32} = 0.100$	0.101 (0.022)	0.025	0.001	0.001	0.970
$\beta_{33} = -0.200$	-0.198 (0.046)	0.049	0.002	0.002	0.962
$\beta_{40} = 2.000$	2.004 (0.049)	0.051	0.004	0.002	0.946
$\beta_{41} = -0.100$	-0.097 (0.016)	0.017	0.003	0.000	0.952
$\beta_{42} = 0.100$	0.100 (0.033)	0.037	0.000	0.001	0.978
$\beta_{43} = -0.200$	-0.205 (0.070)	0.072	-0.005	0.005	0.958
$\beta_{50} = 1.000$	0.891 (0.137)	0.129	-0.109	0.031	0.840
$\beta_{51} = -1.000$	-0.984 (0.042)	0.048	0.016	0.002	0.958
$\beta_{52} = 1.000$	1.034 (0.086)	0.091	0.034	0.009	0.956
$\beta_{53} = -1.000$	-1.035 (0.170)	0.162	-0.035	0.030	0.938
$\sigma_1^2 = 0.160$	0.160 (0.004)	0.004	0.000	0.000	0.982
$\sigma_2^2 = 0.160$	0.160 (0.004)	0.004	0.000	0.000	0.974
$\gamma_1 = 0.250$	0.254 (0.094)	0.099	0.004	0.009	0.940
$\gamma_2 = -0.250$	-0.243 (0.107)	0.114	0.007	0.012	0.960
$\gamma_3 = 0.250$	0.279 (0.137)	0.148	0.029	0.020	0.960
$\gamma_4 = -0.250$	-0.248 (0.085)	0.092	0.002	0.007	0.960
$\gamma_5 = 0.300$	0.264 (0.093)	0.099	-0.036	0.010	0.946
$\zeta = -0.200$	-0.210 (0.175)	0.184	-0.010	0.031	0.962

Table 4.4: Parameter estimates for five-variate simulation scenario. 'Emp. Mean (SD)' denotes the average estimated value with the standard deviation of parameter estimates and Mean SE the mean standard error calculated for each parameter from each model fit. Coverage probabilities are calculated from $\hat{\boldsymbol{\Omega}} \pm 1.96\text{SE}(\hat{\boldsymbol{\Omega}})$. The median [IQR] total computation time (e.g. including time taken to obtain initial estimates etc.) was 27.164 [25.859, 28.991] seconds.

the random effects structure here being an intercept only; in a similar fashion to the binomial previously seen a potential excess of zeroes in the simulated data cloud the amount of variance in the response $\boldsymbol{Y}_i$ attributable to the random effects. The overdispersed generalised Poisson exhibits the largest bias ($\approx -10\%$) for the association parameter $\hat{\gamma}$ with notably less variation than its underdispersed counterpart. The time invariant parameter $\hat{\zeta}$ is well-estimated across all families.

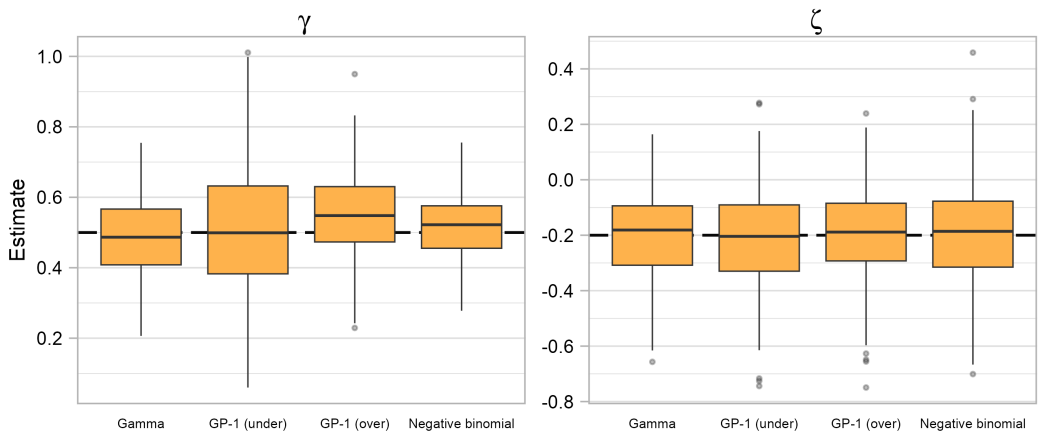


Figure 4.4: Estimates for the survival parameters $\hat{\gamma}$ and $\hat{\zeta}$ for four univariate joint models ('over': overdispersion; 'under': underdispersion). The conditional distribution of $\boldsymbol{Y}_i|\boldsymbol{b}_i$ is shown on the x axis. The dashed line signifies the true parameter value.

Estimates for the scalar dispersion parameter $\hat{\sigma}$ are presented in Figure 4.5. We note here that the Gamma and negative binomial joint models enjoy broadly good estimates of the shape and overdispersion parameter respectively. Both the under- and overdispersed generalised Poisson appear to struggle to capture the true ratio of variance to the mean; markedly worse for the overdispersed scenario. We argue since the estimates $\hat{\boldsymbol{\Omega}}_{-\hat{\boldsymbol{\sigma}}}$ are generally good, and computation time is fast for these families, that this lapse in ability for the algorithm is a trade-off a practitioner may accept; especially since the estimates $\boldsymbol{\sigma}$ are not worlds away from their true values. Here we refer back to the idea of trade-offs between computation time and parameter estimates in Section 3.4.7, with pointers towards usage of subsequent MCMC schemes to obtain 'more accurate' parameter estimates.

Additional results are provided in Appendix B.2.3. Here, we note good performance for the Gamma scenario and more mixed performance for the count responses. For the generalised Poisson, the intercept β_0 is fairly poorly estimated (percentage bias 6%), and σ appears to be overestimated in magnitude for the underdispersed case, as discussed above. For the negative binomial simulations, we note that all fixed effects $\boldsymbol{\beta}$ appear to be fairly poorly estimated, though the remainder of $\hat{\boldsymbol{\Omega}}_{-\hat{\boldsymbol{\beta}}}$ is well-captured including, interestingly, the overdispersion parameter σ. The shortcomings in estimation capabilities for the negative binomial may be due to the magnitude $\boldsymbol{\beta}$ producing simulated $\boldsymbol{Y}_i \in [0, 4890]$ which could elicit struggles in obtaining the true fixed effects coefficients $\boldsymbol{\beta}$. In order to investigate a potential remedy for this, we repeated the simulation with the fixed effects being set as the much smaller $\boldsymbol{\beta} = (0.50, 0.05, 0.10, -0.10)^\top$. Results from this brief set of simulations are additionally presented in Appendix B.2.3, where we note the fixed effects are comparatively well-captured.

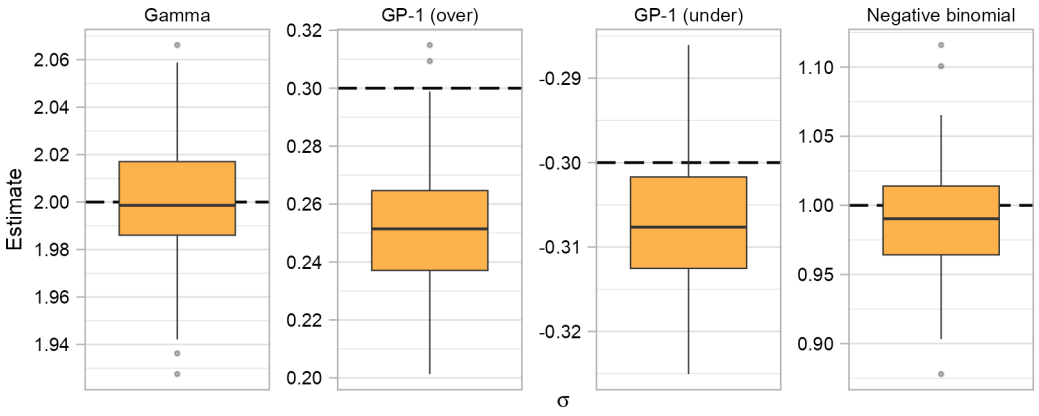


Figure 4.5: Dispersion parameter estimates $\hat{\sigma}$ for four univariate joint models ('over': overdispersion; 'under': underdispersion). The conditional distribution of $\boldsymbol{Y}_i|\boldsymbol{b}_i$ is shown on the panel title. The dashed line signifies the true parameter value.

Finally, appraising the computation times for these families in Table 4.5, we note that both Gamma and GP-1 enjoy computation times not dissimilar from those previously seen. Interestingly, the negative binomial systematically required more iterations (median [IQR] 24 [23, 26]), which coupled with its comparatively more 'expensive' log-likelihood (4.8) lead to these joint model fits being the most computationally cumbersome. We also note that `glmmTMB` (i.e. 'total computation time' in Table 4.5) appears struggle with this family, so this doesn't appear exclusive to the approximate EM algorithm considered. The repeated negative binomial simulations with smaller fixed effects enjoyed lower computation times.

Univariate family	EM algorithm	Total computation time
Gamma	3.561 [3.406, 3.732]	7.188 [6.870, 7.540]
Negative binomial	13.312 [12.507, 14.332]	28.306 [27.379, 29.462]
Generalised Poisson (under)	4.437 [3.687, 5.891]	6.832 [6.005, 8.402]
Generalised Poisson (over)	4.022 [3.660, 4.375]	5.739 [5.303, 6.232]

Table 4.5: Median [IQR] computation times for both the EM algorithm (plus calculation of standard errors) and total computation (i.e. start to finish). All times are given in seconds. The generalised Poisson is split into 'over': overdispersion and 'under': underdispersion as per Table 4.2.

4.5.7 Closing remarks

We have demonstrated in the simulation studies carried out across Sections 4.5.3–4.5.6 that the approximate EM algorithm can estimate the parameters of interest $\boldsymbol{\Omega}$ well in a timely manner, with particular focus on the central tendency for estimates of the survival parameters $\hat{\boldsymbol{\Phi}} = \left(\hat{\boldsymbol{\gamma}}^\top, \hat{\zeta}\right)^\top$ and all estimates $\hat{\boldsymbol{\Omega}}$ provided in Appendix B.2. The implementation here representing a novel approach to estimation under a multivariate joint model

with its longitudinal sub-models composed by GLMMs by maximum likelihood.

We noted broadly good performance in the estimation of $\hat{\boldsymbol{\Omega}}$; however, this is not without its caveats. Poorer coverage was observed for $\hat{\boldsymbol{\Omega}}_{\hat{\mathrm{D}}}$ under the binomial response. We posit that this could be due to the nature of the response itself not carrying as much information i.e. not enough variation to be captured by corresponding GLMM – potentially indicating hard-to-remedy simulation issue – and/or the approximation determined by (3.1)–(3.3) could behave differently under such a binary response, as we go on to investigate in the next chapter. The same interpretation here extends to the poor estimation of the generalised Poisson, with potentially imbalanced/highly variable data leading to struggles in capturing the true parameter values.
In spite of some poorer performance amongst $\hat{\boldsymbol{\Omega}}_{-\hat{\boldsymbol{\Phi}}}$, the coverage of $\hat{\boldsymbol{\Phi}}$ remained respectable irregardless of the conditional distribution imposed on the response.

We argue – as we did in Section 3.4.7 – that there does exist some trade-off between the fast computation time and parameter estimation capabilities. In comparison with the purely Gaussian longitudinal specifications we considered in Section 3.4, the coverage of the fixed effects and variance-covariance matrices appear to suffer slightly, in particular under the binomial GLMM.
We perhaps revisit then the potential use of the approximate EM algorithm in obtaining starting values in a quick manner for some more precise (e.g. MCMC algorithm). The recommendation may then be that practitioners interested in fitting binomial or generalised Poisson should proceed with some caution, or use $\hat{\boldsymbol{\Omega}}$ as $\boldsymbol{\Omega}^{(0)}$ in a subsequent MCMC scheme. All other families, as parameterised in this thesis, are likely to not warrant the same level of caution.

4.6 On the number of quadrature nodes, ϱ

In Section 2.3.4, we briefly alluded to fleeting guidance within joint modelling literature for the proposed amount of quadrature nodes one should employ. Wulfsohn and Tsiatis (1997) utilised $\varrho = 2$ under their, quite simplistic, 'standard' quadrature, with Crowther et al. (2016) using both five and fifteen quadrature points under this methodology, too. Bernhardt et al. (2015) used nine in their adaptive quadrature routine on a multivariate joint model with logistic sub-model.

Rizopoulos (2012a) state that whilst non-adaptive routines (e.g. Section 2.3.2) require, say, 7–15 nodes, adaptive methods (Sections 2.3.3 and 2.3.4) require substantially fewer nodes, say 3–5 to obtain a sufficiently accurate approximation. Our simulation results appear to corroborate this, with three quadrature nodes utilised throughout all simulations carried out in this chapter and Chapter 3.

Furthermore, we could attempt to develop something of a heuristic for the number of quadrature nodes ϱ in a joint model. We return to the recommendation proffered by Stringer and Bilodeau (2022) in a GLMM setting (2.23). This incorporated the sample size n and the minimal observed length of follow-up M. Clearly, it's possible that M could take value one, thereby producing an undefined value for ϱ i.e. in `R log(250, base=1)` returns `Inf`, which is inoperable upon.

We therefore seek to field some potential (and fairly arbitrarily-defined) rules-of-thumb in the joint modelling setting, determining the number of nodes ϱ as a function of some percentile of follow-up length, $\varrho(p)$, and comparing with results obtained using different numbers of quadrature nodes

$$\varrho(p) = \left\lceil \frac{3}{2} \log_{P_p} n - 2 \right\rceil, \tag{4.26}$$

where P_p denotes the p^{th} percentile of lengths of follow-up.

With no generality, we explore $p \in \{10, 25, 33\}$ in addition to candidate values of ϱ previously used in literature $\varrho \in \{2, 3, 5, 7, 9, 15\}$. For the 'base' trivariate joint model simulation outlined in Section 4.5.1, where baseline hazard parameter $\log \nu$ is randomly selected from set of potential values $\{-3, -3.5, -2.75\}$ to yield a variety of failure rates ($\approx$ 20%–40%) for the $N = 200$ simulated sets of data; the idea being that 'static' numbers of nodes *not* dependent on the data may falter slightly.

Figure 4.6 illustrates the root mean square error for each of the parameters whose parameter updates outlined in either Sections 3.3 or 4.4.2 are taken using (adaptive) Gauss-Hermite quadrature. Most parameters conform to a sharp visual levelling-off at $\varrho = 3$, with little distinguishing $\varrho = 3$ and $\varrho = 15$. The residual variance for the Gaussian response appears to fluctuate more in comparison, however we note in this instance from the y-axis that these are in fact minute fluctuations, and could be attributed to MC error. We could perhaps simply summarise these results by concluding that $\varrho = 3$ or $\varrho = 5$ seem reasonable, with diminishing returns for $\varrho > 5$; this result being consistent with earlier comments in Rizopoulos (2012a). The heuristic (4.26) resulted in three to five nodes for both $p = 25$ and $p = 33$ in the majority of simulated sets, but do not appear to out-perform the static choices $\varrho = 3$ or $\varrho = 5$.

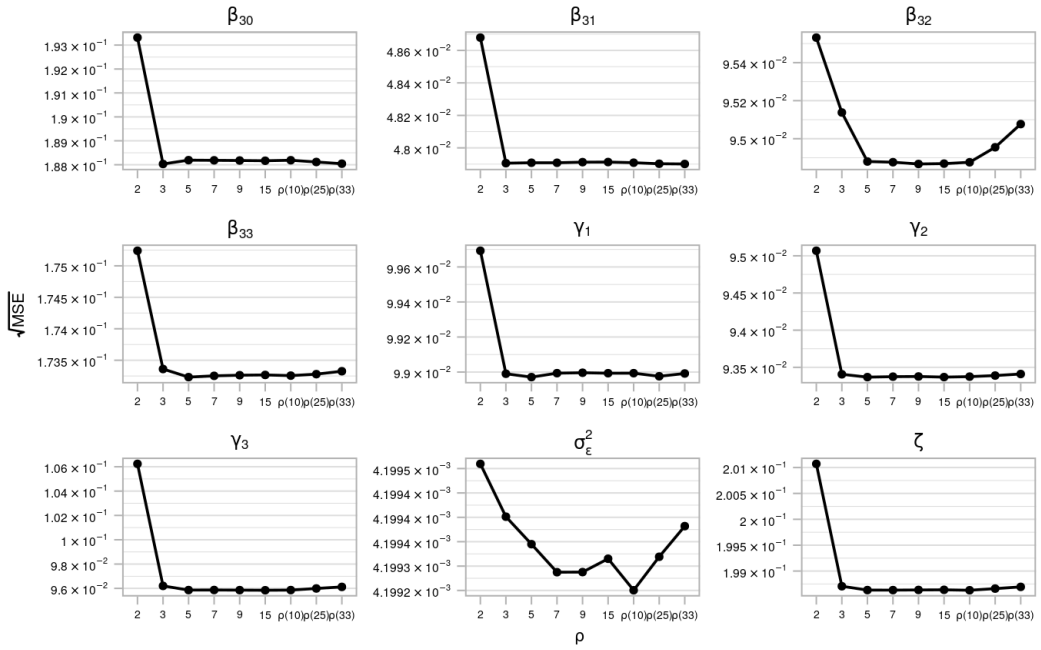


Figure 4.6: Root mean square error (MSE) for a subset of parameter vector $\hat{\boldsymbol{\Omega}}$ for different candidate values $\varrho \in \{2, 3, 5, 7, 9, 15, \varrho(10)\,, \varrho(25)\,, \varrho(33)\}$. Due to `plotmath` restrictions in `R`, ϱ is represented simply by ρ.

4.7 Usage of the normal approximation in a Monte Carlo EM algorithm

In this chapter as well as the last, we have exploited the normal approximation on the conditional distribution of the random effects $f(\boldsymbol{b}_i|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega})$, which we first outlined in Section 3.2.1. This approximation allowed us to 'collapse down' requisite expectations of the form $\mathbb{E}_i[g(\boldsymbol{b}_i)\,|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega}]$ (2.17), which form the potentially computationally taxing E-step, and take them with respect to low-dimensional quadrature. In Section 2.3.5 we demonstrated how this expectation could be appraised using Monte Carlo integration (2.28).

We now seek to combine the normal approximation on the random effects given the observed data and current parameter estimates and the Monte Carlo approach to the E-step outlined. We evaluate the expectation using N Monte Carlo samples

$$\mathbb{E}_i[g(\boldsymbol{b}_i)\,|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega}] \approx \frac{1}{N} \sum_{l=1}^{N} g\Big(\boldsymbol{b}_i^{(l)}\Big)\,, \tag{4.27}$$

where $\boldsymbol{b}_i^{(l)},\ l = 1, \ldots, N$ are drawn from $N\big(\hat{\boldsymbol{b}}_i, \hat{\Sigma}_i\big)$ with $\hat{\boldsymbol{b}}_i$ and $\hat{\Sigma}_i$ calculated from (3.2) and

(3.3), respectively. We seek to investigate differences in EM algorithms wherein the E-step is evaluated by Gauss-Hermite quadrature or Monte Carlo methods (4.27), simulating and fitting one hundred sets of trivariate data with true parameter values outlined in Section 4.5.1.

We appraise the E-step (4.27) at each iteration by antithetic Monte Carlo previously mentioned in Section 2.3.5 and implemented in earlier work (Henderson et al., 2000), as well as a quasi-Monte Carlo approach (Philipson et al., 2020). 'Ordinary' Monte Carlo sampling was not considered as previous work in joint modelling showed the two alternative methods to be more computationally efficient whilst affording the same level of performance (Philipson et al., 2020). Both Monte Carlo approaches to the E-step are implemented as they appear in `joineRML` (Hickey et al., 2018a) and disquisition surrounding the underlying sampling properties of each approach is given in Philipson et al. (2020). The aim of the exercise here is to investigate differences in performance across EM algorithms, all of which make use of the normal approximation on $\boldsymbol{b}_i|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega}$ utilised throughout the thesis.

At the outset of the MCEM algorithm, we set the number of Monte Carlo samples to be taken as $N = 100$; as pointed out by Ripatti et al. (2002) it is likely unwise to draw many samples when $\boldsymbol{\Omega}$ is far away from the maximiser and increase N as we approach these MLEs $\hat{\boldsymbol{\Omega}}$. Using the relative difference criterion in (3.4), say

$$\delta_{\text{rel}}^{(m+1)} = \max\left(\frac{|\Omega_1^{(m+1)} - \Omega_1^{(m)}|}{|\Omega_1^{(m)}| + \nu}, \ldots, \frac{|\Omega_L^{(m+1)} - \Omega_L^{(m)}|}{|\Omega_L^{(m)}| + \nu} \right),$$

Ripatti et al. (2002) define the coefficient of variation ('cv') as

$$\text{cv}\left(\delta_{\text{rel}}^{(m+1)}\right) = \text{Var}\left(\delta_{\text{rel}}^{(m-1)}, \delta_{\text{rel}}^{(m)}, \delta_{\text{rel}}^{(m+1)}\right) \Big/ \text{mean}\left(\delta_{\text{rel}}^{(m-1)}, \delta_{\text{rel}}^{(m)}, \delta_{\text{rel}}^{(m+1)}\right),$$

which informs the choice of N at the next iteration $N \rightarrow N + \lfloor N/5 \rfloor$ if $\text{cv}\left(\delta_{\text{rel}}^{(m+1)}\right) > \text{cv}\left(\delta_{\text{rel}}^{(m)}\right)$. The idea behind this 'automatic' setting of N for each subsequent iteration is, as $\boldsymbol{\Omega}^{(m+1)}$ approaches $\hat{\boldsymbol{\Omega}}$, the executed M step (i.e. the parameter updates) become smaller and are potentially clouded by MC error which we aim to reduce by said increase in the number of samples N. In a slight departure from both Ripatti et al. (2002) and Hickey et al. (2018a), we declare convergence of the MCEM algorithm when the criteria (3.4) is met in two, instead of three, consecutive iterations.

Figure 4.7 illustrates estimation for the survival parameters $\hat{\boldsymbol{\Phi}}$. Across approaches we note broad visual agreement for the parameter estimates across the E-step approaches, except quasi Monte Carlo appears to slightly better estimate the association parameter attached

to the binary response $\hat{\gamma}_3$. The fully tabulated results are presented in Appendix B.2.4, where this trend of agreement across approaches continues, with similar phenomena as noted in e.g. Sections 4.5.3 and 4.5.5 evident across the two Monte Carlo implementations. Interestingly, both MC approaches appear to struggle in their estimation of the Poisson fixed effect intercept $\hat{\beta}_{20}$.

We may conclude from the brief study undertaken here that the normal approximation (3.1) appears to extend well to an approximate Monte Carlo EM algorithm (or perhaps rather a Monte Carlo approximate EM algorithm) which could be more appealing to some audiences; essentially, the approximation is not limited to quadrature outlined in Sections 3.3, 4.4.2 and 4.4.3

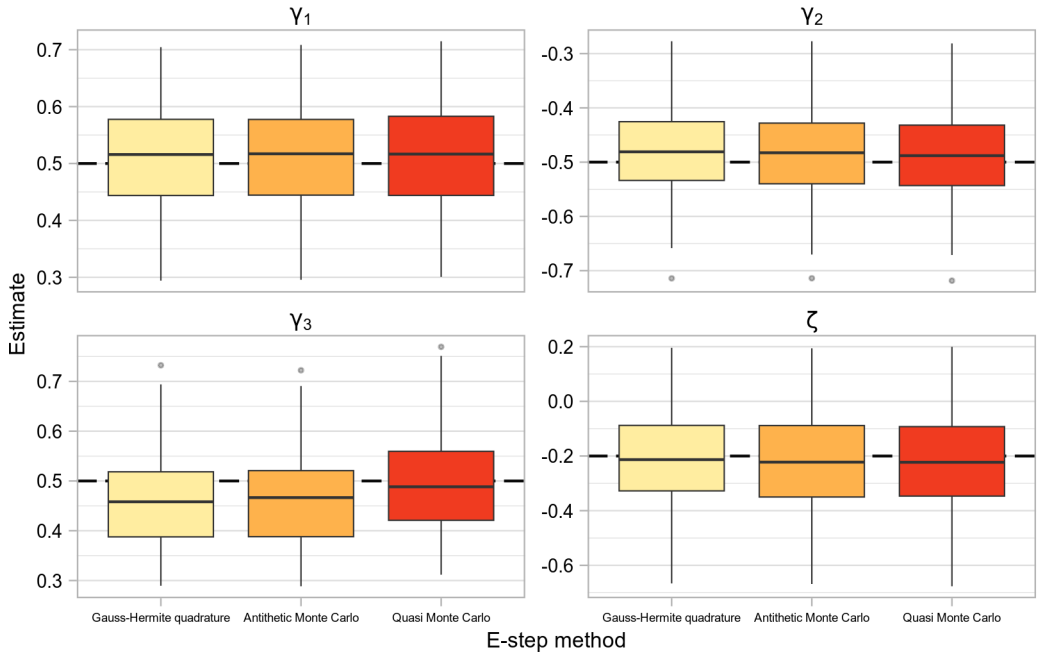


Figure 4.7: Estimates for survival parameters $\hat{\boldsymbol{\gamma}}$ and $\hat{\zeta}$ for a trivariate mixture joint model fit by the approximate EM algorithm as outlined in Chapters 3 and 4 as well as by a Monte Carlo EM algorithm as outlined in Section 4.7. The dashed line signifies the true parameter value.

Chapter 5

Justification for the Normal Approximation

5.1 Approximation foundations and simulation objectives

5.1.1 Background

We seek to justify usage of the normal approximation as we laid out in Section 3.2.1. As mentioned, Rizopoulos (2012a) previously noted – in the context of a univariate (Gaussian) joint model – that the conditional (log) density of the random effects $\log f(\boldsymbol{b}_i|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega})$ is dominated by the (log) density of the linear mixed effects model $\log f(\boldsymbol{Y}_i|\boldsymbol{b}_i; \boldsymbol{\Omega})$, and resembles a (multivariate) normal distribution. Specifically Rizopoulos (2012a) present under certain regularity conditions as $m_i \to \infty$,

$$\boldsymbol{b}_i|T_i, \Delta_i, \boldsymbol{Y}_i; \boldsymbol{\Omega} \overset{\text{appx.}}{\sim} N\big(\tilde{\boldsymbol{b}}_i, \tilde{\Sigma}_i\big), \tag{5.1}$$

with $\tilde{\boldsymbol{b}}_i = \underset{\boldsymbol{b}_i}{\text{argmax}} \log f(\boldsymbol{b}_i, \boldsymbol{Y}_i|\boldsymbol{\Omega})$, and the variance of the estimate $\tilde{\Sigma}_i$ defined similarly to (3.3). Note we discussed this approximation under the guise of numerical integration in Section 2.3.4, but have restated it here for completeness' sake.

Rizopoulos (2012a) utilise their approximation on the linear mixed effects model in the context of pseudo-adaptive Gauss-Hermite quadrature, such that their abscissae and weights are only calculated *once* for each subject; centered at each subject's posterior mode $\tilde{\boldsymbol{b}}_i$ and scaled by the Cholesky factor of $\tilde{\Sigma}_i$. In contrast, we utilise weights and abscissae based on a $N(0, 1)$ kernel and scale these based on $\hat{\boldsymbol{b}}_i$ and $\hat{\Sigma}_i$ at *each* EM iteration in order to approximate requisite conditional expectations as shown in Section 3.3.

Bernhardt et al. (2015) extended the approximation (5.1) to a joint model with a logistic sub-model in place of the survival (3.1); later extended to the 'classic' joint models we considered in Chapter 2 by Murray and Philipson (2022); and finally the emergent joint models with (at least one) GLMM sub-model (Murray and Philipson, 2023).

5.1.2 Simulation objectives

Consider a subject with the set of observed (univariate) data $\mathcal{D}_i = \{\boldsymbol{Y}_i, T_i, \Delta_i\}$, which was generated under the known set of 'true' parameters

$$\boldsymbol{\Omega}^{(\text{TRUE})} = \left(\boldsymbol{\beta}^{(\text{TRUE})\top}, \boldsymbol{\sigma}^{(\text{TRUE})\top}, \gamma^{(\text{TRUE})}, \zeta^{(\text{TRUE})}\right)^\top$$

following the simulation strategy in e.g. Section 3.4.1. We want to contend that the posterior distribution of their random effects, $f\left(\boldsymbol{b}_i | \mathcal{D}_i; \boldsymbol{\Omega}^{(\text{TRUE})}\right)$ approximately follows a (multivariate) normal distribution centered on $\hat{\boldsymbol{b}}_i$ with curvature $\hat{\Sigma}_i$. That is, the normal approximation on the distribution of the random effects conditioned on the observed data and parameter estimates laid out in Section 3.2.1 and used throughout the thesis is reasonable.

In addition, earlier results (Baghishani and Mohammadzadeh, 2012; Rizopoulos, 2012a) suggest that an approximation such as (5.1) should improve as an increased longitudinal profile length is observed i.e. $m_i \to \infty$, which we aim to verify. Simultaneously, we further want to argue that the result from Baghishani and Mohammadzadeh (2012) – that the posterior distribution of random effects for *any* GLMM are asymptotically normal – extends to the "generalised" joint modelling scenario (4.4). Lastly, we undertake investigation into the differences *between* previously-used approximations from Rizopoulos (2012a), i.e. $N\left(\tilde{\boldsymbol{b}}_i, \tilde{\Sigma}_i\right)$ (5.1) and Bernhardt et al. (2015), i.e. $N\left(\hat{\boldsymbol{b}}_i, \hat{\Sigma}_i\right)$ (3.1).

5.1.3 Simulation strategy

With our objectives delineated, we now outline the simulation strategy taken to achieve them.

For $\boldsymbol{\Omega}^{(\text{TRUE})}$ determined by the 'standard' simulation scenarios specified in Sections 3.4.1 and 4.5.1 for the Gaussian and non-Gaussian cases, respectively, we simulate *one* set of data, $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_n\}$, using the methodology outlined in Section 2.5. In a slight departure from Section 4.5.1, the binomial and (overdispersed) generalised Poisson are instead simulated under an intercept-and-slope with $\text{vech}(\text{D}) = (0.50, 0.125, 0.09)^\top$. The Gompertz baseline hazard with scale $\log \nu = -3$ and shape $\alpha = 0.1$ produces a relatively low failure rate, allowing for an abundance of profiles of maximal length $m_i = 10$, whilst

simultaneously resulting in a '*grading*' of longitudinal profile lengths $m_i = 1, \ldots, 9$, which should provide insight into the suitability of the approximation as m_i gradually grows large.

Remark. We set $r = 10$ in keeping with our 'default' simulation scenario in Section 3.4.1. The approximation – ostensibly improving as $m_i \to \infty$ – unsurprisingly benefits from a longer longitudinal profile given its underlying assumption of normality. We *don't* consider larger r in what we believe is a neat compromise between a 'realistic' length of follow-up and a visual argument for the approximation's improvement with m_i.

We proceed then as follows, for each $i = 1, \ldots, n$:

1. Calculate $\hat{\boldsymbol{b}}_i$ and $\hat{\Sigma}_i$ by (3.2) and (3.3), respectively, with $\boldsymbol{\Omega}^{(m)}$ substituted by $\boldsymbol{\Omega}^{(\text{TRUE})}$ and $\boldsymbol{\lambda}_0^{(m)}(\cdot)$ by $\exp\{\log\nu + \alpha\boldsymbol{u}^\top\}$ where $\boldsymbol{u}$ is the vector of failure times in the simulated set of data $\mathcal{D}$. If no survival part is required, then we instead calculate $\tilde{\boldsymbol{b}}_i$ and $\tilde{\Sigma}_i$ by methodology in Section 5.1.1.

2. Simulate from $f\big(\boldsymbol{b}_i|\mathcal{D}_i;\boldsymbol{\Omega}^{(\text{TRUE})}\big)$ using a Metropolis-Hastings scheme with 1,000 iterations of burn-in and 10,000 iterations thereafter. This results in two vectors of correlated 'draws' from the target density: $\boldsymbol{b}_{i0}^{(\text{MH})}$ and $\boldsymbol{b}_{i1}^{(\text{MH})}$. We obtain proposals for the random effects from the multivariate t distribution with 4 degrees of freedom, $t_4\big(\hat{\boldsymbol{b}}_i, a\hat{\Sigma}_i\big)$, where a is a scalar tuning parameter controlled to ensure efficient sampling from the target distribution. This Metropolis-Hastings acceptance rate is captured in A_i.

3. Generate 10,000 samples from $N\big(\hat{\boldsymbol{b}}_i, \hat{\Sigma}_i\big)$, or indeed $N\big(\tilde{\boldsymbol{b}}_i, \tilde{\Sigma}_i\big)$ if required, using e.g. `rmvnorm` (Genz et al., 2021) and produce a data ellipse based on this sample (Fox and Weisberg, 2019; Friendly et al., 2013): $\mathcal{E}$, signifying where one expects the majority of the 'datapoints' (i.e. the sample determined by the approximate distribution) to lie at the 95% confidence level.

4. Calculate the proportion of the generated samples from $f\big(\boldsymbol{b}_i|\mathcal{D}_i;\boldsymbol{\Omega}^{(\text{TRUE})}\big)$ which lie within the region bounded by $\mathcal{E}$, denoting this quantity ψ_i. The methodology used in determining this is given in Appendix A.7.

The above routine then equips us with the posterior density $f\big(\boldsymbol{b}_i|\mathcal{D}_i;\boldsymbol{\Omega}^{(\text{TRUE})}\big)$; the posterior modes and (co)variances, say $\big\{\hat{\boldsymbol{b}}_i, \hat{\Sigma}_i\big\}$, as well as the analogous quantities calculated under omission of $f\big(T_i, \Delta_i|\boldsymbol{b}_i;\boldsymbol{\Omega}^{(\text{TRUE})}\big)$, $\big\{\tilde{\boldsymbol{b}}_i, \tilde{\Sigma}_i\big\}$, $i = 1, \ldots, n$.

We then visualise the 'true' posterior $f\big(\boldsymbol{b}_i|\mathcal{D}_i;\boldsymbol{\Omega}^{(\text{TRUE})}\big)$ with $\mathcal{E}$, showing where the 'mass' of the theoretical distribution *should* lie, superimposed; investigate the relationship between ψ_i and m_i; and explore differences in the normal distributions $N\big(\hat{\boldsymbol{b}}_i, \hat{\Sigma}_i\big)$ and $N\big(\tilde{\boldsymbol{b}}_i, \tilde{\Sigma}_i\big)$.

5.2 Results

5.2.1 Visualisations

We begin by undertaking purely visual appraisal of the approximation obtained by the strategy outlined in Section 5.1.3.

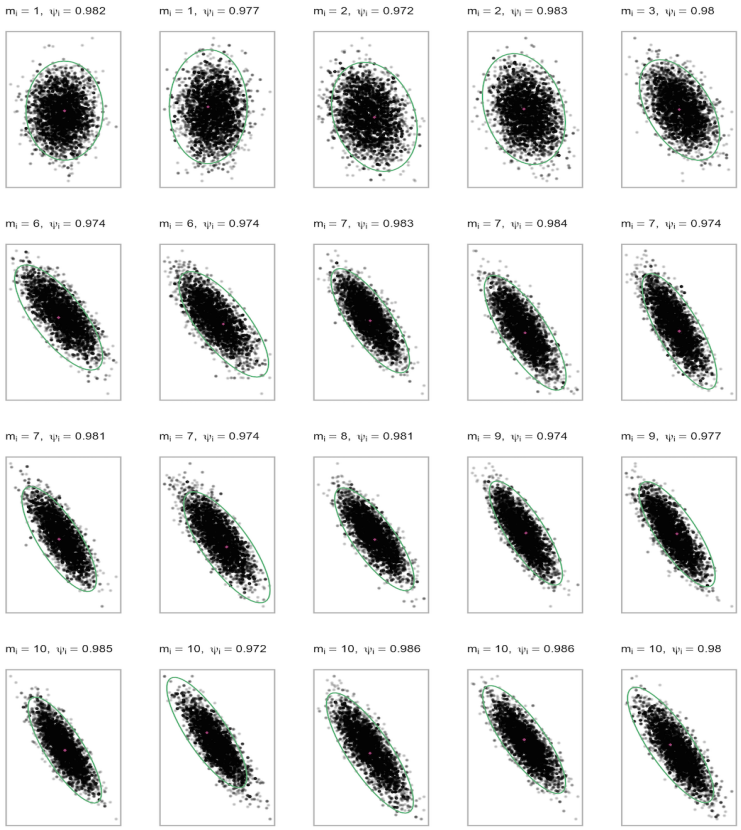


Figure 5.1: Scatterplot of the sample $f\left(\boldsymbol{b}_i|\mathcal{D}_i;\boldsymbol{\Omega}^{(\text{TRUE})}\right)$ with overlaid ellipse showing the theoretical distribution $N\left(\hat{\boldsymbol{b}}_i,\hat{\Sigma}_i\right)$ for $\boldsymbol{Y}_i|\boldsymbol{b}_i;\boldsymbol{\Omega}^{(\text{TRUE})} \sim N(\cdot,\cdot)$. The 'full' set of visualisations are provided in Appendix C.2.1. The random slopes are on the y-axis and intercepts on the x

We ensure only subjects for whom $f\left(\boldsymbol{b}_i|\mathcal{D}_i;\boldsymbol{\Omega}^{(\text{TRUE})}\right)$ has been efficiently sampled from (i.e. neither 'stuck' in one place, nor freely exploring the real line) by employing the common heuristic of only considering subjects with $0.20 < A_i < 0.25$.

We select 40 simulated subjects to produce such visualisations for, and present these on a per-family basis in Appendix C.2.1. When deciding which profiles to show, we *always* attempt to show the 'graded' incrementing of m_i and its effect on the feasibility of the approximation. This is achieved by first ordering the simulated subjects by m_i, further ordering within this follow-up length by their (purely arbitrary) allocated subject identifier and taking the first 40 resultant samples; subsequently applying the process delineated in

Section 5.1.3.

On the whole, for each family we observe that the samples from the posterior density $f\left(\boldsymbol{b}_i|\mathcal{D}_i;\boldsymbol{\Omega}^{(\text{TRUE})}\right)$ lie comfortably within the theoretical $\mathcal{E}$. We observe *generally* that the scatter of sampled random effects becomes increasingly concentrated as m_i grows large, which is reflected by commensurate shrinkage exhibited in $\mathcal{E}$. Indeed, as m_i increases, there will be less uncertainty in the estimate for $\hat{\boldsymbol{b}}_i$, which is reflected in the resultant distribution $N\left(\hat{\boldsymbol{b}}_i,\hat{\Sigma}_i\right)$, which generates said ellipse.

However, this is by no means steadfast. For instance, when the response of interest is a count there are noticeable fluctuations in the spread, even for the longest profiles m_i. This could be due to the magnitude of the response here leading to greater uncertainty in the estimate for $\hat{\boldsymbol{b}}_i$ (i.e. a larger resultant covariance matrix $\hat{\Sigma}_i$). Additionally, say for the binomial family, there is some notable 'shifting' occurring, resulting in the sample from $f\left(\boldsymbol{b}_i|\mathcal{D}_i;\boldsymbol{\Omega}^{(\text{TRUE})}\right)$ not best 'lining up' with $\mathcal{E}$. In spite of this, there is still a good proportion of overlap between the true distribution of $\boldsymbol{b}_i|\mathcal{D}_i;\boldsymbol{\Omega}^{(\text{TRUE})}$ and $N\left(\hat{\boldsymbol{b}}_i,\hat{\Sigma}_i\right)$. To demonstrate this and allow for ease of immediate comparison, Figures 5.1 and 5.2 present a subset of these visualisations for the Gaussian and binomial cases, respectively; the latter exhibiting said 'shifting' and the former generally conforming.

To bring this purely visual inspection exercise to a close, we perhaps conclude that – for the profiles shown in Appendix C.2.1 – the approximation appears to be reasonable, with there generally being tangible (visual) benefit from the presence of a longer period of follow-up.

5.2.2 Relationship between m_i and ψ_i

Next, we undertake a slightly different simulation strategy: Increasing the number of subjects to $n = 1000$, and repeating the process solely monitoring m_i, A_i, and ψ_i.

In a frequentist sense, if the normal approximation is reasonable for the density $f\left(\boldsymbol{b}_i|\mathcal{D}_i;\boldsymbol{\Omega}^{(\text{TRUE})}\right)$, we would expect approximately 95% of the samples from this posterior to lie *within* the region bounded by $\mathcal{E}$. Due to the stochastic nature of Metropolis-Hastings scheme utilised, or other reasons to be discussed, there may be some variability in this proportion. Given we obtain 10^4 samples from both $f\left(\boldsymbol{b}_i|\mathcal{D}_i;\boldsymbol{\Omega}^{(\text{TRUE})}\right)$ and $N\left(\hat{\boldsymbol{b}}_i,\hat{\Sigma}_i\right)$, we can perhaps be confident such a contribution would be relatively small, with observed discrepancies then inherent to the approximation itself.

We note from Figure 5.3 that there is evidence to suggest that $N\left(\hat{\boldsymbol{b}}_i,\hat{\Sigma}_i\right)$ is over-confident, as we appear to routinely overshoot the nominal 0.95 we expect from theory. As previously ascribed, we don't expect this to be due to variability in the sampling process, but perhaps could attribute to some other behaviour of $N\left(\hat{\boldsymbol{b}}_i,\hat{\Sigma}_i\right)$. For instance, this may produce a

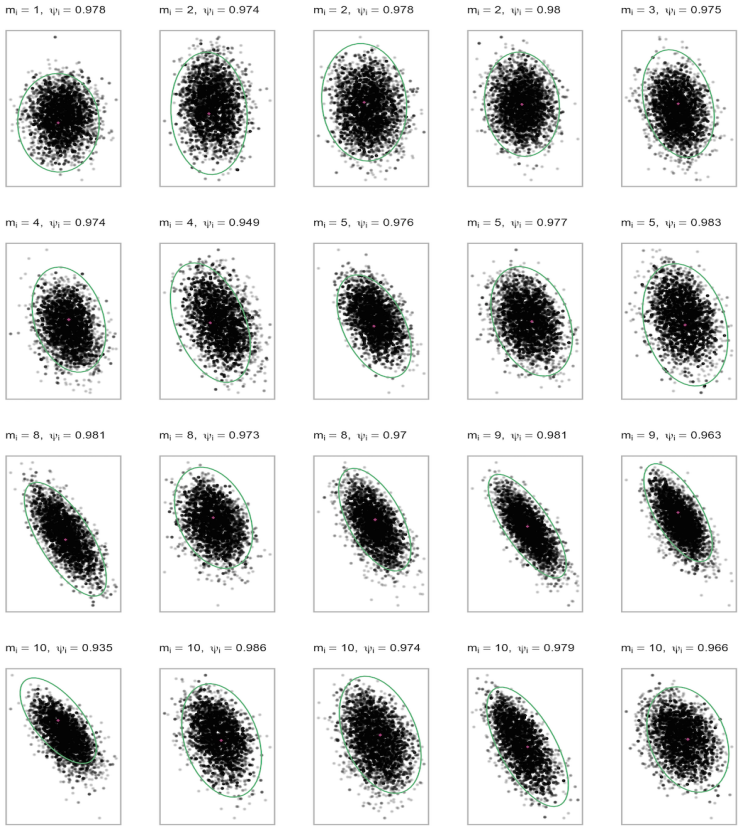


Figure 5.2: Scatterplot of the sample $f\left(\boldsymbol{b}_i|\mathcal{D}_i;\boldsymbol{\Omega}^{(\text{TRUE})}\right)$ with overlaid ellipse showing the theoretical distribution $N\left(\hat{\boldsymbol{b}}_i,\hat{\Sigma}_i\right)$ for $\boldsymbol{Y}_i|\boldsymbol{b}_i;\boldsymbol{\Omega}^{(\text{TRUE})} \sim \text{Bin}(\cdot)$. The 'full' set of visualisations are provided in Appendix C.2.1. The random slopes are on the y-axis and intercepts on the x

density more variable than $f\left(\boldsymbol{b}_i|\mathcal{D}_i;\boldsymbol{\Omega}^{(\text{TRUE})}\right)$; exhibit strange behaviour in the tails; or otherwise be less/more concentrated about the modal mass, such that the additional ≈ 0.025–0.030 coverage we observe could be accounted for.

In order to investigate the over-confidence noted in Figure 5.3, we briefly turn attention to the produced densities themselves. We follow earlier work (Murray and Philipson, 2023) and 'zoom in' on the densities for the random effects for randomly-chosen subjects who fall into different 'halves' of follow-up, $m_i \in [1,5]$ and $m_i \in [6,10]$, presenting the posterior $f\left(\boldsymbol{b}_i|\mathcal{D}_i;\boldsymbol{\Omega}^{(\text{TRUE})}\right)$ with the theoretical $N\left(\hat{\boldsymbol{b}}_i,\hat{\Sigma}_i\right)$ overlaid in Figure 5.4. Here we observe some behaviour which could account for the additional coverage we noted in Figure 5.3: The normal distribution sometimes fully encapsulates the notably more concentrated posterior distribution i.e. $\hat{\Sigma}_i$ in $N\left(\hat{\boldsymbol{b}}_i,\hat{\Sigma}_i\right)$ *overestimates the variance*. Interestingly this occurs more frequently around the modal mass of $f\left(\boldsymbol{b}_i|\mathcal{D}_i;\boldsymbol{\Omega}^{(\text{TRUE})}\right)$ rather than in its tails, which one may expect a priori.

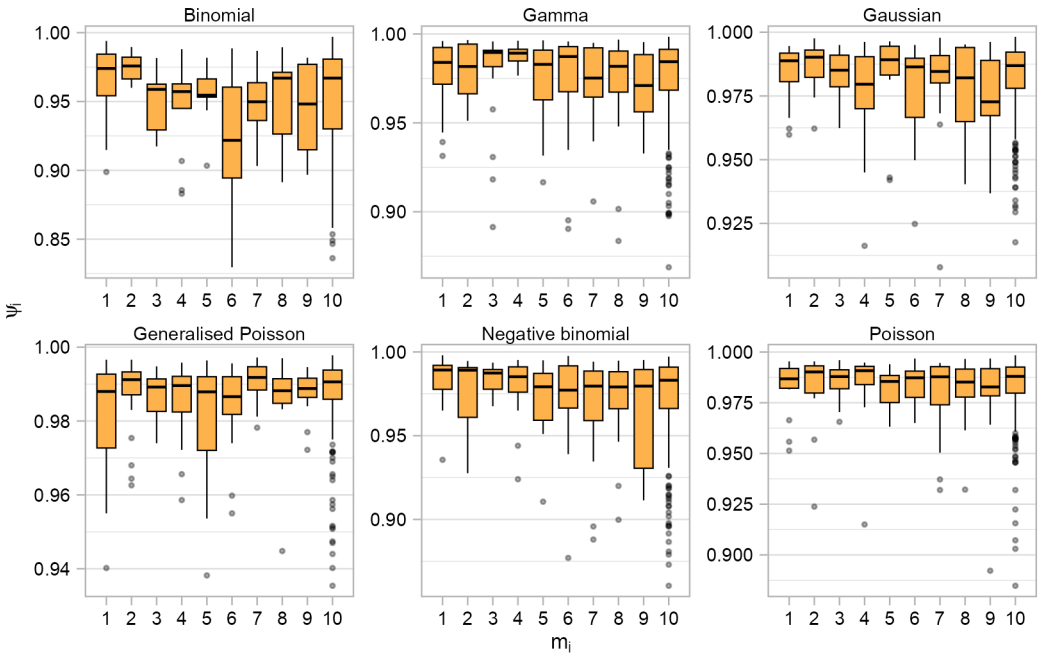


Figure 5.3: Boxplots of ψ_i against graded profile lengths m_i for each family considered in Chapters 2–4

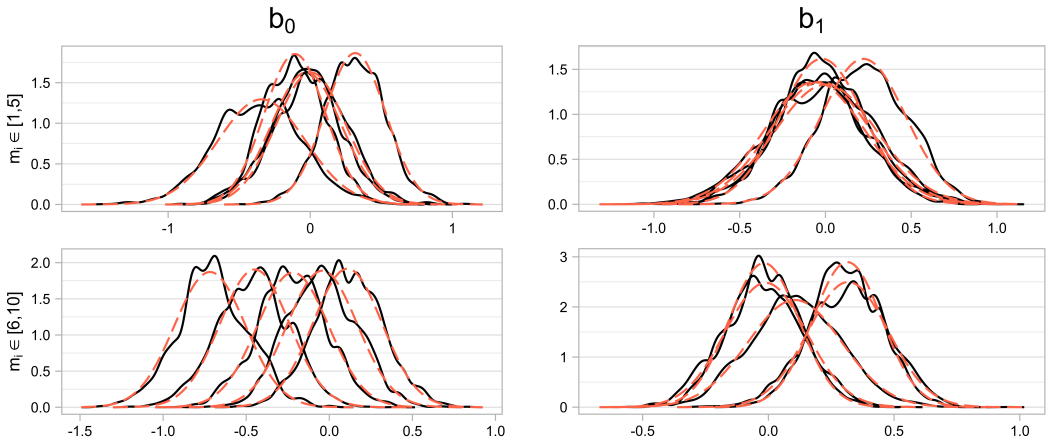


Figure 5.4: Posterior distributions for $b_{i0}^{(\mathrm{MH})}$ and $b_{i1}^{(\mathrm{MH})}$ for five randomly chosen subjects who fall into two follow-up 'halves'. The solid black line is the sampled posterior density and the red dashed line shows $N\big(\hat{\boldsymbol{b}}_i, \hat{\Sigma}_i\big)$.

5.2.3 Effect of the inclusion of the survival density

Thus far we have conducted an investigation into the normal approximation $N\big(\hat{\boldsymbol{b}}_i, \hat{\Sigma}_i\big)$, with focus on the ellipse $\mathcal{E}$ arising from a large number of draws from this distribution. We now turn our attention to the characteristics of $\mathcal{E}$ itself; namely on the differences in normal distributions centered and scaled by $\hat{\boldsymbol{b}}_i$ and $\hat{\Sigma}_i$ (i.e. Section 3.2) in contrast with $\tilde{\boldsymbol{b}}_i$

and $\tilde{\Sigma}_i$ (Section 5.1.1). That is, monitoring the effect of including $f\left(T_i, \Delta_i | \boldsymbol{b}_i; \boldsymbol{\Omega}^{(\text{TRUE})}\right)$ in calculating the location and curvature of resultant (multivariate) normal distributions.

We look to achieve this by studying the components of $\mathcal{E}$ generated by the modal estimate $\hat{\boldsymbol{b}}_i$ (and its covariance $\hat{\Sigma}_i$), and $\tilde{\boldsymbol{b}}_i$ ($\tilde{\Sigma}_i$) calculated by accounting for the survival part ('S') in the complete data log-likelihood, and without ('NS' i.e. 'no survival'), respectively. For $n = 1000$ simulated subjects we calculate $\left\{\hat{\boldsymbol{b}}_i, \hat{\Sigma}_i\right\}$ and $\left\{\tilde{\boldsymbol{b}}_i, \tilde{\Sigma}_i\right\}$ and subsequently construct the ellipses $\mathcal{E}^{(S)}$ and $\mathcal{E}^{(NS)}$, respectively. We then compare: Estimates for $\hat{\boldsymbol{b}}_i$ and $\tilde{\boldsymbol{b}}_i$ (dictating the origin of the ellipse); the semi-minor, $r_y^{(\text{S})}$ and $r_y^{(\text{NS})}$, and semi-major axes, $r_x^{(\text{S})}$ and $r_x^{(\text{NS})}$. The interested reader is pointed to Appendix A.7 for information relating to the derivation of these quantities.

For each family, we present the difference in the random intercept and slope, say $\hat{\boldsymbol{b}}_i - \tilde{\boldsymbol{b}}_i$ as well as the difference in the semi-minor $r_y^{(\text{S})} - r_y^{(\text{NS})}$ and semi-major axes $r_x^{(\text{S})} - r_x^{(\text{NS})}$. The quantities related to the ellipsis' axes directly stem from the estimated $\hat{\Sigma}_i$, with smaller variances leading to smaller minor and major axes. In addition to the presence of $f\left(T_i, \Delta_i | \boldsymbol{b}_i; \boldsymbol{\Omega}^{(\text{TRUE})}\right)$, we hold secondary interest in the effect an increased profile length m_i has on the difference observed in these quantities.

Negative values for the difference in (semi-)minor and major axes implies the ellipse containing 95% of the theoretical data sample is *larger* when the survival density is *not* included i.e. $\hat{\Sigma}_i$ is less variable than $\tilde{\Sigma}_i$. Differences in modal estimates $\hat{\boldsymbol{b}}_i - \tilde{\boldsymbol{b}}_i$ simply represents the modal 'mass' lying at a different point.

The Gaussian case is presented in Figure 5.5 and the binomial in Figure 5.6. In order to reduce bloat, all remaining families are presented in Appendix C.2.2. For the Gaussian case, estimates for the intercept and slope are both larger *without* inclusion of the survival density in their calculation, particularly for larger m_i; implying incorporation of survival information, correlated with the longitudinal process, influences the optimisation routine for the value $\hat{\boldsymbol{b}}_i$ compared to $\tilde{\boldsymbol{b}}_i$. The ellipsis' axes are always slightly smaller in both the x- and y-direction under inclusion of the survival process, implying that the variance is larger with its removal. This perhaps makes sense given the additional information afforded by inclusion of the time-to-event process. There exists some visual banding here between lengths of follow-up: At lower m_i there's little difference in the semi-major axis, which increases with increased m_i.

In the binomial case, we note a more apparent difference in random effects estimates, with $\hat{\boldsymbol{b}}_i$ being smaller in magnitude than $\tilde{\boldsymbol{b}}_i$, indicating that the exclusion of the survival information leads to larger estimates from the optimisation routine with comparatively larger variances attached, too.

These results perhaps suggests for the binomial specifically, referring back to results from

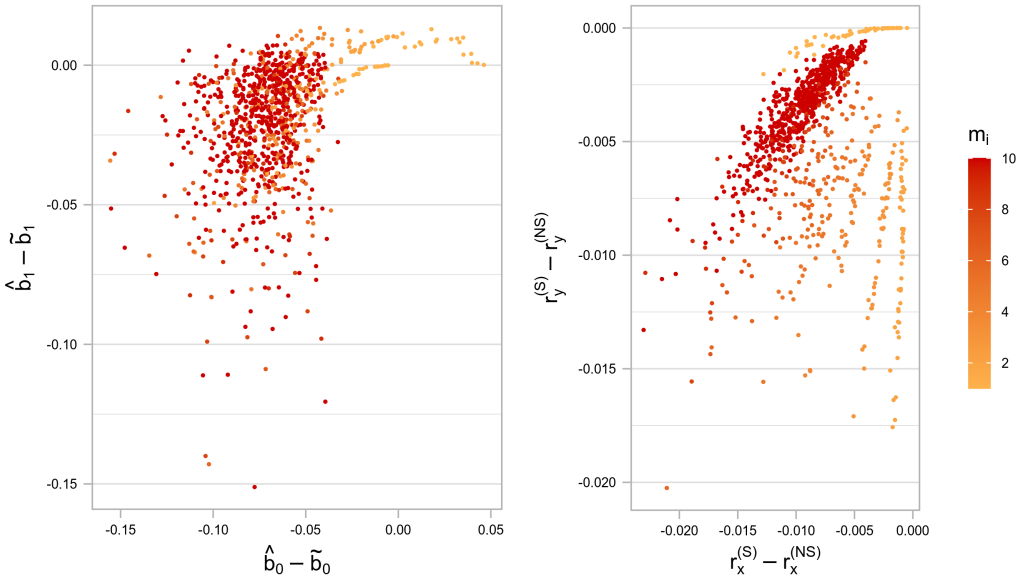


Figure 5.5: Difference in the modal estimates $\hat{\boldsymbol{b}}_i - \tilde{\boldsymbol{b}}_i$; semi-minor axis $r_y^{(\mathrm{S})} - r_y^{(\mathrm{NS})}$; and semi-major $r_x^{(\mathrm{S})} - r_x^{(\mathrm{NS})}$. The differences themselves arise from the *removal* of the survival density from the complete data log-likelihood in the process to obtain the modal estimate and its covariance. The differences themselves arise from the *removal* of the survival density from the complete data log-likelihood in the process to obtain the modal estimate and its covariance for $\boldsymbol{Y}_i|\boldsymbol{b}_i \sim N(\cdot, \cdot)$.

Rizopoulos (2012a) discussed in Section 5.1.1, that the joint density $f(\boldsymbol{b}_i, \boldsymbol{Y}_i, T_i, \Delta_i; \boldsymbol{\Omega})$ here is *not* dominated by the density $f(\boldsymbol{Y}_i|\boldsymbol{b}_i; \boldsymbol{\Omega})$. Taking this idea further, we undertake a brief simulation study and consider the Gaussian, Poisson and binomial families. Using the median sample values from $f\big(\boldsymbol{b}_i|\mathcal{D}_i; \boldsymbol{\Omega}^{(\mathrm{TRUE})}\big)$ we evaluate the proportion of the log-likelihood accounted for by $\log f\big(T_i, \Delta_i|\boldsymbol{b}_i; \boldsymbol{\Omega}^{(\mathrm{TRUE})}\big)$ across each $i = 1, \ldots, n$ simulated subjects. The median [IQR] proportion obtained for the Gaussian was 12.06 [6.23, 22.25]%, for the Poisson 2.75 [2.15, 3.88]%, and the binomial 11.21 [6.44, 32.18]%. Perhaps here the wider range observed in the binomial has a knock-on effect on the subsequent accuracy $N\big(\hat{\boldsymbol{b}}_i, \hat{\Sigma}_i\big)$.

To briefly summarise the differences in modal estimates and quantities relating to the ellipse's axes for the other exponential families presented in Appendix C.2.2, we observe that both the intercept and slope estimates are larger if the survival density is *excluded*. For the negative binomial, Gamma and Poisson case, these observed differences are relatively small in magnitude; the generalised Poisson exhibiting a larger difference.

Interestingly, the estimated variance (i.e. $\hat{\Sigma}_i$ against $\tilde{\Sigma}_i$) is larger with the survival density included for the generalised Poisson and Poisson, and smaller for all other families. We note however that the differences here are all relatively close to zero, save for the binomial case once more, where the difference is more dramatic.

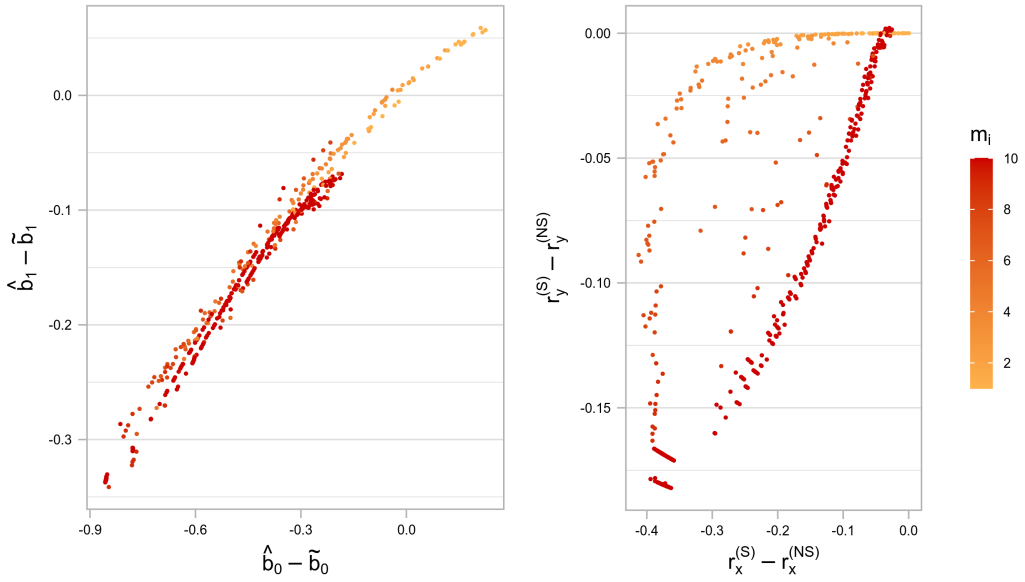


Figure 5.6: Difference in the modal estimates $\hat{\boldsymbol{b}}_i - \tilde{\boldsymbol{b}}_i$; semi-minor axis $r_y^{(\mathrm{S})} - r_y^{(\mathrm{NS})}$; and semi-major $r_x^{(\mathrm{S})} - r_x^{(\mathrm{NS})}$. The differences themselves arise from the *removal* of the survival density from the complete data log-likelihood in the process to obtain the modal estimate and its covariance. The differences themselves arise from the *removal* of the survival density from the complete data log-likelihood in the process to obtain the modal estimate and its covariance for $\boldsymbol{Y}_i|\boldsymbol{b}_i \sim \mathrm{Bin}(\cdot)$.

5.3 Concluding remarks

Bringing this exercise to a close, we review our objectives detailed in Section 5.1.2 and aim to provide concise conclusionary statements. We approached all objectives by sampling from the 'true' posterior $f\left(\boldsymbol{b}_i|\mathcal{D}_i;\boldsymbol{\Omega}^{(\mathrm{TRUE})}\right)$ and comparing to the theoretical distribution arising from the normal approximation both with, $N\left(\hat{\boldsymbol{b}}_i, \hat{\Sigma}_i\right)$ (Bernhardt et al., 2015; Murray and Philipson, 2022, 2023), and without, $N\left(\tilde{\boldsymbol{b}}_i, \tilde{\Sigma}_i\right)$ (Rizopoulos, 2012a), the survival density considered in order to evaluate some of our objectives.

Primarily, we wanted to argue that the normal approximation on the conditional distribution of the random effects appears reasonable. We initially carried out a purely visual approach in Section 5.2.1, wherein we posited that for all presented profiles, there visually was a good deal of agreement.

In Section 5.2.2 we sought to *quantify* the agreement by summarising the quantity ψ_i for each candidate exponential distribution considered in Chapters 3 and 4. Here we noted that the coverage given by $N\left(\hat{\boldsymbol{b}}_i, \hat{\Sigma}_i\right)$ was over-confident, providing approximately 2.5-3% extra coverage than the nominal. We concluded here that the (co)variance $\hat{\Sigma}_i$ is over-estimated. This finding, which was consistent across families, could lead to development of some scale factor applied to $\hat{\Sigma}_i$ in order to generate a distribution truer to

$f\big(\boldsymbol{b}_i|\mathcal{D}_i;\boldsymbol{\Omega}^{(\text{TRUE})}\big)$. These cross-family findings appeared to cement earlier results from Baghishani and Mohammadzadeh (2012).
Finally, we investigated the approximated distribution itself in Section 5.2.3, wherein we investigated more thoroughly the difference in approximations by comparing resultant estimates for $\hat{\boldsymbol{b}}_i$ and components relating to the ellipse (representing said approximated distribution). We noted changes which weren't overtly substantial; corroborating earlier results from Rizopoulos (2012a).

To provide something of a denouement, we conclude here that the approximation *is* reasonable. However, we note these simulations are non-exhaustive: The results here appear to present the approximation in a fairly positive manner, but there may be scenarios where the approximation is not suitable e.g. due to some data characteristics. Indeed, as we lamented in Section 2.5, there exist *many* tuning knobs when simulating data under a joint model; throughout Section 5.2 we presented only *one* set of parameters $\boldsymbol{\Omega}^{(\text{TRUE})}$. Future work indeed could involve monitoring the effect of certain parameters, or other data characteristics as we performed in earlier simulations (e.g. Section 3.4). Lastly, we only considered intercept-and-slope parameterisations of the random effects – largely to allow for simpler visualisations – where in actuality different specifications may be pertinent.

5.4 On the shape of required integrands

When fitting joint models, the approximate EM algorithm is used to evaluate expectations of the form $\mathbb{E}_i[g(\boldsymbol{b}_i)\,|T_i,\Delta_i,\boldsymbol{Y}_i;\boldsymbol{\Omega}]$ as outlined in Sections 3.3, 4.4.2, and 4.4.3. We exploited the normal approximation (3.1) to evaluate these against $N\big(\hat{\boldsymbol{b}}_i,\hat{\Sigma}_i\big)$ instead of 'full' conditional distribution $f(\boldsymbol{b}_i|T_i,\Delta_i,\boldsymbol{Y}_i;\boldsymbol{\Omega})$ in (2.16).

For these expectations, we seek to provide the posterior distribution of the term in the corresponding integrand i.e. $g(\boldsymbol{b}_i)\,|T_i,\Delta_i,\boldsymbol{Y}_i;\boldsymbol{\Omega}$ as means of justification for appraisal via adaptive Gauss-Hermite quadrature (since the resultant distribution resembles the normal) or at the median value of the log-normal distribution in instances where $g(\boldsymbol{b}_i)\,|\cdot = \exp\Big\{\mathrm{X}_i\boldsymbol{\beta}+\mathrm{Z}_i\hat{\boldsymbol{b}}_i\Big\}$. We therefore proceed by considering these disparate methods of integral evaluation.

We acquire these posteriors by first fitting a joint model to a set of simulated data to obtain MLEs $\hat{\boldsymbol{\Omega}}$. Next, for the *first* simulated subject who survived follow-up, we sample from $f\big(\boldsymbol{b}_i|T_i,\Delta_i,\boldsymbol{Y}_i;\hat{\boldsymbol{\Omega}}\big)$ 50,000 times after 1,000 iterations of burn-in using a Metropolis-Hastings scheme in order to obtain $\boldsymbol{b}_i^{(\text{MH})}$ and then appraising the appropriate function $g\big(\boldsymbol{b}_i^{(\text{MH})}\big)$ at *each* of these draws; the acceptance rate of the sampling scheme is controlled to be between 23-28%. The resultant posterior distribution of $g(\boldsymbol{b}_i)\,|T_i,\Delta_i,\boldsymbol{Y}_i;\hat{\boldsymbol{\Omega}}$ is then shown for this subject, allowing for visual inspection at (the arbitrarily chosen) start,

middle, and end of follow-up.

5.4.1 Expectations evaluated by Gauss-Hermite quadrature

Across Chapters 3 and 4 we require

$$\mathbb{E}_i\left[\exp\left\{\mathrm{S}_i\boldsymbol{\zeta}+\sum_{k=1}^{K}\gamma_k\mathrm{F}_{\boldsymbol{u}_ik}\boldsymbol{b}_{ik}\right\}\middle|T_i,\Delta_i,\boldsymbol{Y}_i;\boldsymbol{\Omega}\right], \tag{5.2}$$

for updates to both the baseline hazard $\lambda_0(\cdot)$ and survival parameters $\boldsymbol{\Phi}$. Additionally, in Chapter 4 specifically when updating the parameter vector $\boldsymbol{\Omega}^{(m)}\rightarrow\boldsymbol{\Omega}^{(m+1)}$ under a binomial-, negative binomial-, or generalised Poisson-distributed response we needed to evaluate

$$\begin{aligned}&\mathbb{E}_i[\log(\mathbf{1}+\exp\{\mathrm{X}_i\boldsymbol{\beta}+\mathrm{Z}_i\boldsymbol{b}_i\})\,|T_i,\Delta_i,\boldsymbol{Y}_i;\boldsymbol{\Omega}],\\ &\mathbb{E}_i[\log(\boldsymbol{\varphi}_i+\exp\{\mathrm{X}_i\boldsymbol{\beta}+\mathrm{Z}_i\boldsymbol{b}_i\})\,|T_i,\Delta_i,\boldsymbol{Y}_i;\boldsymbol{\Omega}],\quad\text{and}\\ &\mathbb{E}_i[\log(\boldsymbol{Y}_i\boldsymbol{\varphi}_i+\exp\{\mathrm{X}_i\boldsymbol{\beta}+\mathrm{Z}_i\boldsymbol{b}_i\})\,|T_i,\Delta_i,\boldsymbol{Y}_i;\boldsymbol{\Omega}],\end{aligned} \tag{5.3}$$

respectively.

The posterior density of $g(\boldsymbol{b}_i)\,|T_i,\Delta_i,\boldsymbol{Y}_i;\hat{\boldsymbol{\Omega}}$ for the survival expectation (5.2) is shown in Figure 5.7 which we note adequately resembles the normal distribution across failure times and so our implementation using adaptive Gauss-Hermite quadrature is reinforced here. The same is noted for the quantity $g(\boldsymbol{b}_i)$ housed in the expectation related to the negative binomial, binomial, and generalised Poisson cases are presented in Figure 5.8, Appendix C.2.3, and Appendix C.2.4, respectively. We note for the binomial case presented in the Appendix that the resultant density has slightly heavier tails than one may expect a normally distributed quantity to have; perhaps explaining some of the poorer performance noted for the binomial case in Chapter 4.

5.4.2 Expectations evaluated at median value of the log-normal

In Chapter 4 we frequently required the expectation

$$\mathbb{E}_i[\exp\{\mathrm{X}_i\boldsymbol{\beta}+\mathrm{Z}_i\boldsymbol{b}_i\}\,|T_i,\Delta_i,\boldsymbol{Y}_i;\boldsymbol{\Omega}] \tag{5.4}$$

in the E-step for GLMMs. Given the normal approximation we utilise as outlined in Section 3.2.1 – and specifically (3.8) – the quantity (5.4) is log-normally distributed, i.e.

$$\begin{aligned}\exp\{\boldsymbol{\eta}_i\}&\overset{\text{appx.}}{\sim}LN(\hat{\boldsymbol{\mu}}_i,\mathrm{A}_i)\qquad\text{since}\\ \boldsymbol{\eta}_i&\overset{\text{appx.}}{\sim}N(\hat{\boldsymbol{\mu}}_i,\mathrm{A}_i),\end{aligned}$$

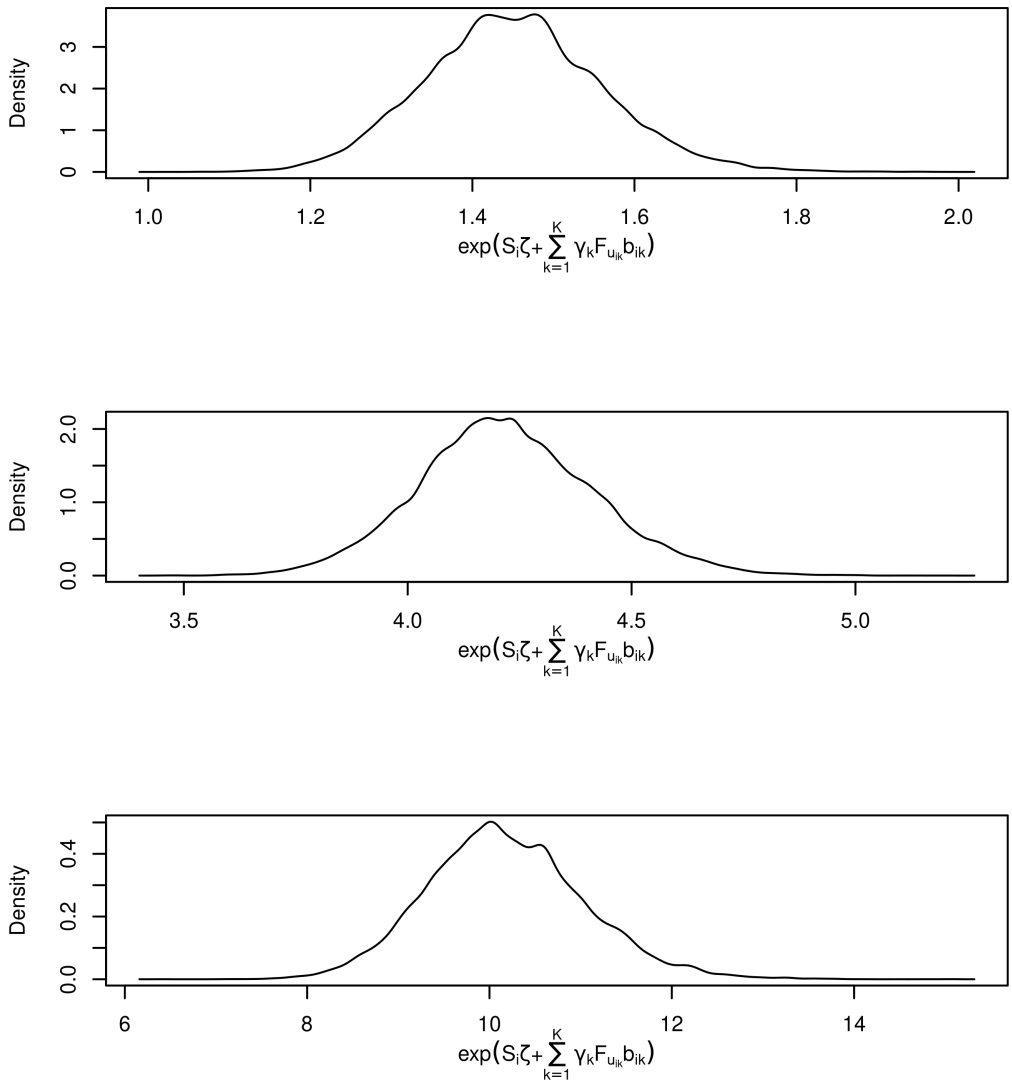


Figure 5.7: Posterior density for $g(\boldsymbol{b}_i)\,|T_i, \Delta_i, \boldsymbol{Y}_i; \hat{\boldsymbol{\Omega}} = \exp\Big\{\mathrm{S}_i\boldsymbol{\zeta} + \sum_{k=1}^{K}\gamma_k \mathrm{F}_{\boldsymbol{u}_i k}\boldsymbol{b}_{ik}\Big\}$ evaluated at the first (top pane), middle and final (bottom pane) failure time for a set of simulated bivariate Gaussian data with $\boldsymbol{\gamma}^{(\mathrm{TRUE})} = (0.5, -0.5)^{\top}$.

where 'LN' denotes 'log-normal'. The mean and median of the log-normally distributed random variable $X \sim LN\big(\mu, \sigma^2\big)$ is, respectively, $\exp\Big\{\mu + \frac{\sigma^2}{2}\Big\}$ and $\exp\{\mu\}$. Given our (multivariate) specification above and previously in (3.8), these are respectively

$$\exp\left\{\hat{\boldsymbol{\mu}}_i + \frac{\boldsymbol{\tau}_i^2}{2}\right\}, \qquad \exp\{\hat{\boldsymbol{\mu}}_i\}\,. \tag{5.5}$$

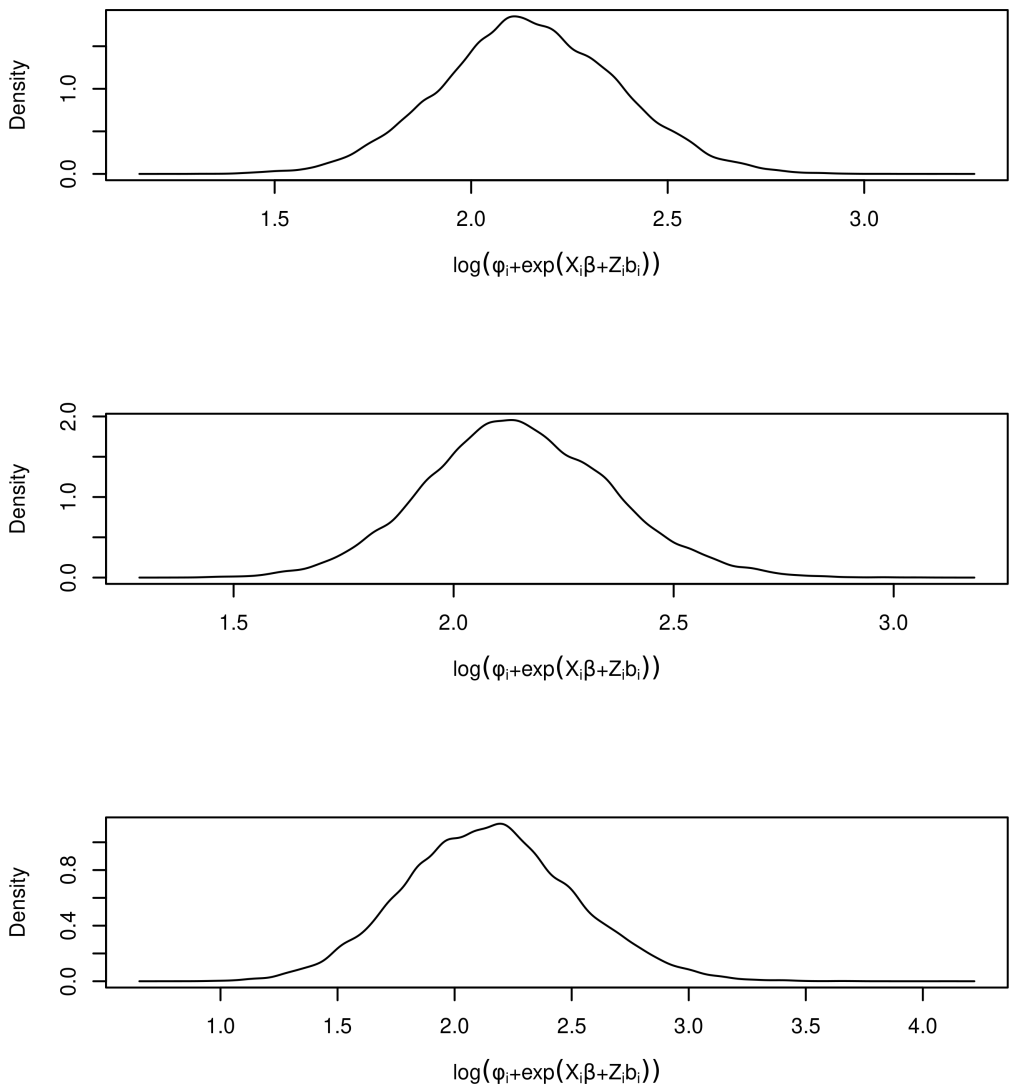


Figure 5.8: Posterior density for $g(\boldsymbol{b}_i)\,|T_i, \Delta_i, \boldsymbol{Y}_i; \hat{\boldsymbol{\Omega}} = \log(\boldsymbol{\varphi}_i + \exp\{\mathrm{X}_i\boldsymbol{\beta} + \mathrm{Z}_i\boldsymbol{b}_i\})$ evaluated at the first (top pane), middle and final (bottom pane) follow-up time for univariate negative binomial simulated data.

Throughout the parameter updates we presented in Chapter 4 for the fixed effects and dispersion parameters in Sections 4.4.2 and 4.4.3 we evaluated the expectation (5.4) at the median value in (5.5) with reasoning that it is more centered at the modal mass of the posterior distribution of $g(\boldsymbol{b}_i)\,|T_i, \Delta_i; \boldsymbol{\Omega} = \exp\{\boldsymbol{\eta}_i\}$, which we seek to justify here.

We present the posterior distribution for the Poisson GLMM sub-model in Figure 5.9. Interestingly, at the start of follow-up ($t_1 = 0$) the distribution resembles a slightly right

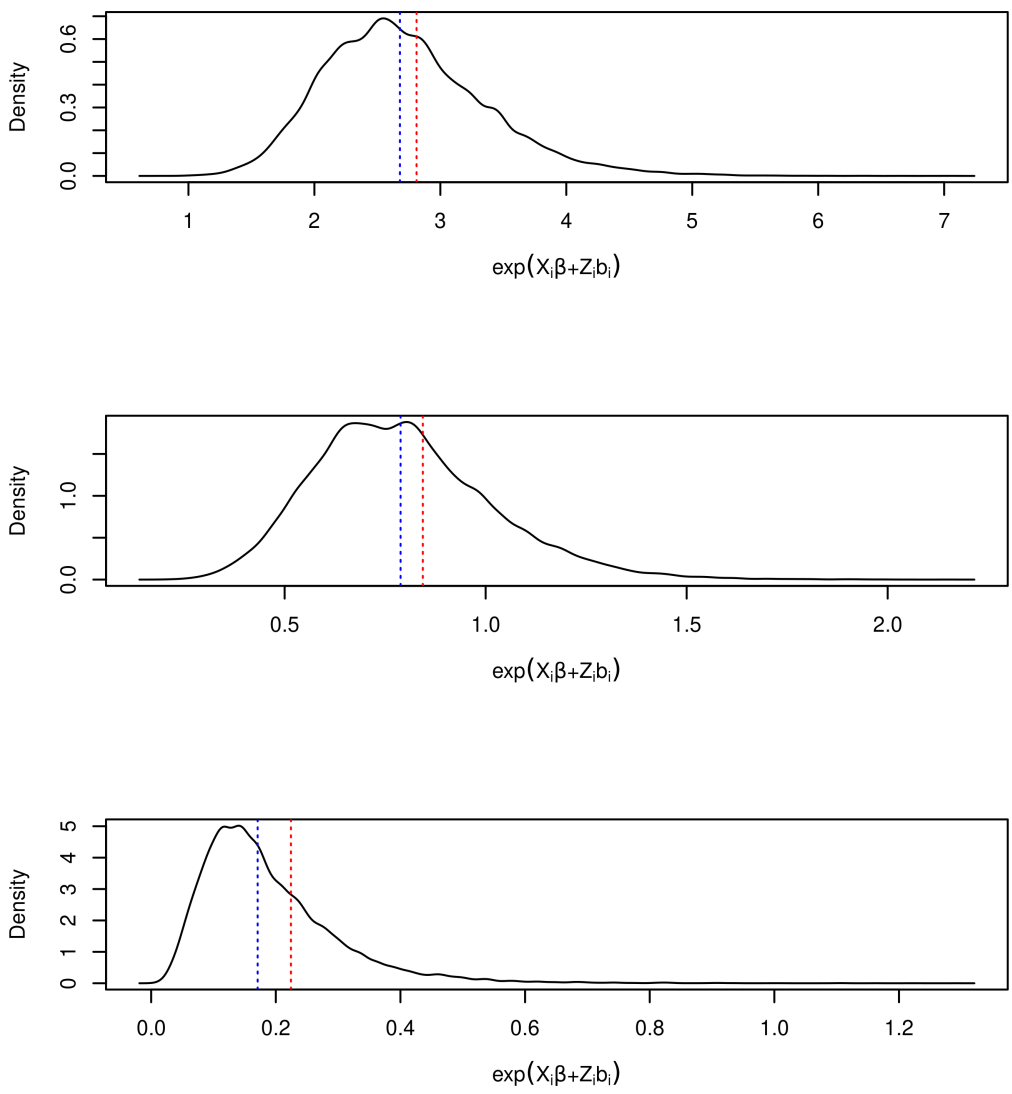


Figure 5.9: Posterior density for $g(\boldsymbol{b}_i)\,|T_i, \Delta_i, \boldsymbol{Y}_i; \hat{\boldsymbol{\Omega}} = \exp\{\mathrm{X}_i\boldsymbol{\beta} + \mathrm{Z}_i\boldsymbol{b}_i\}$ evaluated at the first (top pane), middle and final (bottom pane) follow-up time for univariate Poisson simulated data. The mean and median of the posterior distribution are denoted by the red and blue dashed lines, respectively.

skew normal. The skewness i.e. log-normality is much more marked in later stages of follow-up. We note that for each panel in Figure 5.9 the median is located in the modal mass of the posterior, with the mean being quite clearly more inappropriate at greater levels of skewness.

The same distribution is presented for the binomial, Gamma, negative binomial, and

generalised Poisson case in Appendix C.2.5 where we once more note that the median appears more reasonable, with distributions gradually exhibiting more skew. Additionally, we present the modal value for the Poisson case in Appendix C.2.6; here we note that the mode of the log-normal, $\exp\{\hat{\boldsymbol{\mu}}_i - \boldsymbol{\tau}_i^2\}$, appears to 'undershoot' the peak of the posterior distribution of interest. This, along with the mean 'overshooting' the modal mass, perhaps cements our earlier finding that the normal approximation overestimates the variance in $\hat{\Sigma}_i$ on average.

5.5 Scaling $\hat{\Sigma}_i$ to achieve nominal coverage

Despite being reasonably satisfied with the normal distribution $N(\hat{\boldsymbol{b}}_i, \hat{\Sigma}_i)$, we collected some evidence in Sections 5.2.2 and 5.4.2 that the variance term $\hat{\Sigma}_i$ (3.3) is *over*estimating the true variability in $f(\boldsymbol{b}_i|\mathcal{D}_i; \boldsymbol{\Omega}^{(\text{TRUE})})$. Indeed, in Section 5.3 we hypothesised a scale factor may be developed and applied to $\hat{\Sigma}_i$ such that $N(\hat{\boldsymbol{b}}_i, \hat{\Sigma}_i)$ more closely resembles $f(\boldsymbol{b}_i|\mathcal{D}_i; \boldsymbol{\Omega}^{(\text{TRUE})})$. That is, ψ_i is closer to the nominal 0.95.

Denoting the coverage of the posterior density $f(\boldsymbol{b}_i|\mathcal{D}_i; \boldsymbol{\Omega}^{(\text{TRUE})})$ given the bivariate normal distribution with parameters $\hat{\boldsymbol{b}}_i$ and $\hat{\Sigma}_i$ by $\psi_i(\hat{\boldsymbol{b}}_i, a\hat{\Sigma}_i)$ where $a > 0$ is a scale factor, we consider the minimisation problem of the objective

$$Q(a) = \frac{1}{n}\sum_{i=1}^{n} \left|\psi_i(\hat{\boldsymbol{b}}_i, a\hat{\Sigma}_i) - 0.95\right| \tag{5.6}$$

across $n = 100$ simulated subjects. Since this is one-dimensional in a, we calculate $\min_a Q(a)$ one hundred times using `optim` with Brent's algorithm in the constrained search space for $a \in (0.01, 2.00)$.

	a	
Family	Median [IQR]	Minimum, maximum
Gaussian	0.75 [0.74, 0.77]	0.71, 0.83
Gamma	0.77 [0.76, 0.79]	0.71, 0.85
Poisson	0.77 [0.75, 0.78]	0.72, 0.83
Negative binomial	0.79 [0.77, 0.81]	0.73, 0.85
Generalised Poisson (over)	0.77 [0.75, 0.79]	0.70, 0.86
Binomial	0.76 [0.75, 0.78]	0.71, 0.82

Table 5.1: Median [IQR] along with minimum and maximum values for a which minimise $Q(a)$ in (5.6) across one hundred sets of data simulated under the family shown.

Owing to computational cost, only 5,000 posterior samples are taken from $f(\boldsymbol{b}_i|\mathcal{D}_i; \boldsymbol{\Omega}^{(\text{TRUE})})$ for each $i = 1, \ldots, 100$. Table 5.1 lists the median [IQR], along with the minimum and maximum, estimates for a obtained under previously described 'stock' simulation scenarios for

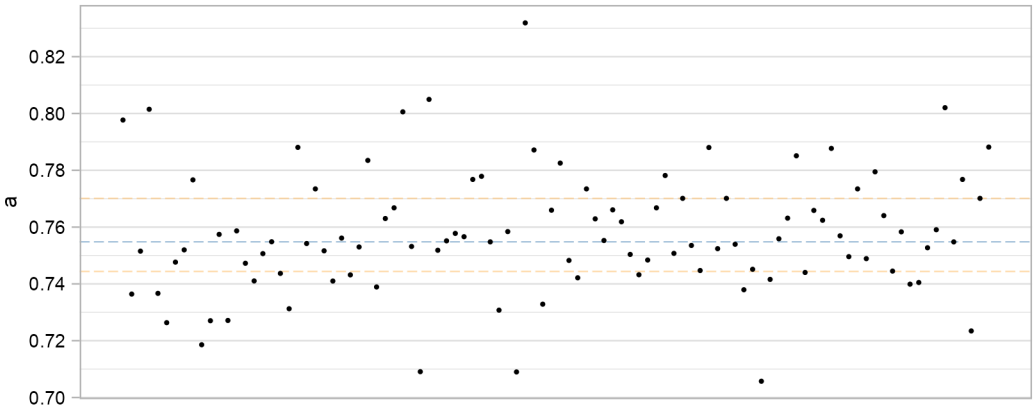


Figure 5.10: Values for a which minimise $Q(a)$ in (5.6); each point represents the value a from *one* set of Gaussian simulated data. The median value is denoted by the blue dashed line and the orange dashed lines represent the interquartile range.

families considered in Chapter 4. Here, the binomial and (overdispersed) generalised Poisson are instead simulated under an intercept-and-slope with $\text{vech}(\text{D}) = (0.50, 0.125, 0.09)^\top$. We note that in all cases the scale factor a reduces the (co)variance present in $\hat{\Sigma}_i$ obtained by (3.3), indicating that overestimation does occur. The results for the Gaussian case are presented in Figure 5.10 where we observe median value $a \approx 0.75$ achieves nominal coverage. Visualisations in a similar spirit to Figure 5.10 are provided in Appendix C.2.7 for the other families considered.

The investigation carried out here is non-exhaustive and very much represents a line of future enquiry. We note in Table 5.1 that there are small differences between families, which could indicate necessity of a scale factor *matrix* in the multivariate case i.e. scaling along the block diagonal of $\hat{\Sigma}_i$. Indeed, we considered only the univariate case and thus not potential impact of applying the scalar a to covariance *between* random effects. The approximate EM algorithm as implemented in Chapters 3 and 4 appears to perform well without any such scaling of the covariance $\hat{\Sigma}_i$; this scale factor a would need to be applied 'during' the EM algorithm itself.

Chapter 6

Post-hoc Analyses, Prediction, and Prognostic Accuracy

With a joint model fit, we turn can now our attention to validation of the model across multiple streams. In this chapter, we consider the residuals of the constituent parts of a fitted joint model; how one may undertake a model-building exercise using long-established testing approaches; and introduce prognostic prediction obtained from a joint model along with arising measures of predictive accuracy. The idea being that they will inform decision-making in the next chapter. For illustration purposes we consider 'dummy'/'toy' examples from appropriate models fit to the PBC data first introduced in Section 1.3.

6.1 Residuals

Residuals are an important byproduct of any modelling process; playing a role in model assessment, diagnostics and improvement. In a joint model, we obtain residuals which are attached to each of the K longitudinal processes, as well as the event-time sub-model i.e. no special 'joint' residual exists, to the author's knowledge.

6.1.1 Residuals for the longitudinal sub-model(s)

Remaining firmly couched within the context of the GLMM sub-model specification which we considered in Chapter 4, we are interested in obtaining conditional subject-specific residuals, for the k^{th} response modelled for subject i at time-point j

$$\begin{aligned} \hat{r}_{ikj} &= Y_{ikj} - \hat{Y}_{ikj}, \\ &= Y_{ikj} - h_k^{-1}\left(\boldsymbol{X}_{ikj}^{\top}\hat{\boldsymbol{\beta}}_k + \boldsymbol{Z}_{ikj}^{\top}\hat{\boldsymbol{b}}_{ik}\right), \end{aligned} \tag{6.1}$$

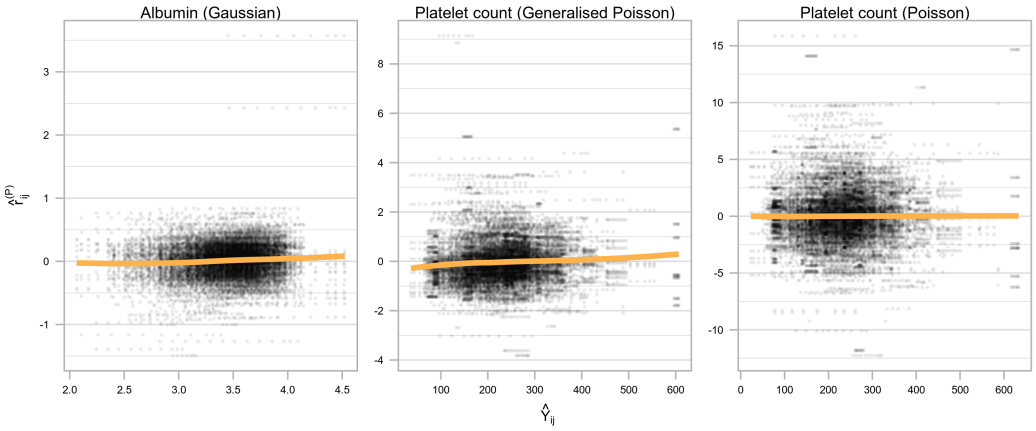


Figure 6.1: Plots of (standardised) residuals $\hat{\boldsymbol{r}}^{(P)}$ against fitted values $\hat{\boldsymbol{Y}}$ for three *separate* GLMMs fit with `glmmTMB`. An orange LOESS line is super-imposed to better reveal underlying trends.

where the quantity being subtracted is the predicted value at the MLEs $\hat{\boldsymbol{\Omega}}$, which notably takes the estimate for $\hat{\boldsymbol{b}}_i$ determined at $\hat{\boldsymbol{\Omega}}$, i.e. (3.2) with $\boldsymbol{\Omega}$ substituted by $\hat{\boldsymbol{\Omega}}$. Since we entertain multiple distributions for $\boldsymbol{Y}_{ik}|\boldsymbol{b}_{ik}$, we operate largely under the standardised (or 'Pearson') residuals,

$$\hat{r}_{ikj}^{(P)} = \frac{\hat{r}_{ikj}}{\sqrt{\text{Var}\left[\hat{Y}_{ikj}\right]}}, \tag{6.2}$$

where $\hat{r}_{ikj}$ is given in (6.1) and the variance $\text{Var}[\cdot]$ was given on a per-distribution basis in Section 4.3.3 along with chosen link function $h_k(\cdot)$ in Table 4.1.

These residuals can then be used to quickly ascertain how well the fitted joint model captures the longitudinal responses it incorporates. That is, we can judge post-hoc whether the residuals for the k^{th} response, $\hat{\boldsymbol{r}}_k^{(P)} = \left(\hat{\boldsymbol{r}}_{1k}^{(P)\top}, \ldots, \hat{\boldsymbol{r}}_{nk}^{(P)\top}\right)^\top$, satisfy the usual assumptions of e.g. homoscedasticity across all subjects. As an example, Figure 6.1 shows the resultant standardised residuals against fitted value plots from three *separate* GLMM fits to albumin (Gaussian) and platelet count (Poisson and generalised Poisson), each fit with a random intercept and slope with fixed effects of a drug-time interaction. Albumin looks to aptly conform to the usual assumptions. Under the Poisson distribution platelet counts appears to violate the modelling assumptions more seriously than under the alternative generalised Poisson.

6.1.2 Residuals for the survival sub-model

We can produce residuals for the time-to-event process modelled by the survival sub-model; providing insight into how well the joint model, i.e. the longitudinal process(es) in addition to the baseline covariates modelled, captures underlying patterns of survival in the observed data.

Many residuals have been proposed for the Cox PH model constructing the survival sub-model, and the interested reader is referred to Chapter 4 in Therneau and Grambsch (2000). Since such discussions are outside of the scope of the presented work, and introduce concepts not considered elsewhere in the thesis, we opt instead to continue with sole consideration of *one* residual, namely the Cox-Snell residual (Cox and Snell, 1968)

$$\hat{r}_i^{(CS)} = \int_0^{T_i} \hat{\lambda}_0(u) \exp\left\{ \boldsymbol{S}_i^\top \hat{\boldsymbol{\zeta}} + \sum_{k=1}^{K} \hat{\gamma}_k \boldsymbol{W}_k(u)^\top \hat{\boldsymbol{b}}_{ik} \right\} du, \tag{6.3}$$

which represents the cumulative hazard under the fitted model for each subject evaluated at their observed failure time. This notably includes the estimate for the baseline hazard which we detailed in Section 3.3.4 and $\hat{\boldsymbol{b}}_i$ is calculated at the MLEs $\hat{\boldsymbol{\Omega}}$ in the same manner outlined in the previous section.

To assess the goodness of fit of the residuals obtained in (6.3), the Kaplan-Meier estimate of their survival function obtained by the `survfit` function from the `R` package `survival` (Therneau, 2015) is compared against the theoretical survival function under the null hypothesis, $S_0(t) = \exp\{-\Lambda_0(t)\}$, with evaluations at observed failure times $S_0(T_i)\,,\ i = 1, \ldots, n$; Rizopoulos (2012b) detail that this is the unit exponential distribution.

As an example, we consider two Cox proportional hazards models fit to the PBC data, one which models the hazard of mortality by recipiency of the study drug, and another by the subject's sex. The produced graphical representations are presented in Figure 6.2. A sign of a well-fitting model is for the Kaplan-Meier estimate to roughly follow the superimposed unit exponential. For the two univariate Cox PH models fit here, we note that although both appear to generally fit the data well, there is some lack of fit for the model only considering the subject's sex, particularly for higher values of $\hat{r}_i^{(CS)}$. This phenomena is not replicated in the 'drug' pane, where the survival function of the Cox-Snell residuals follow this theoretical distribution more snugly, lying within the 95% confidence band for the survival function.

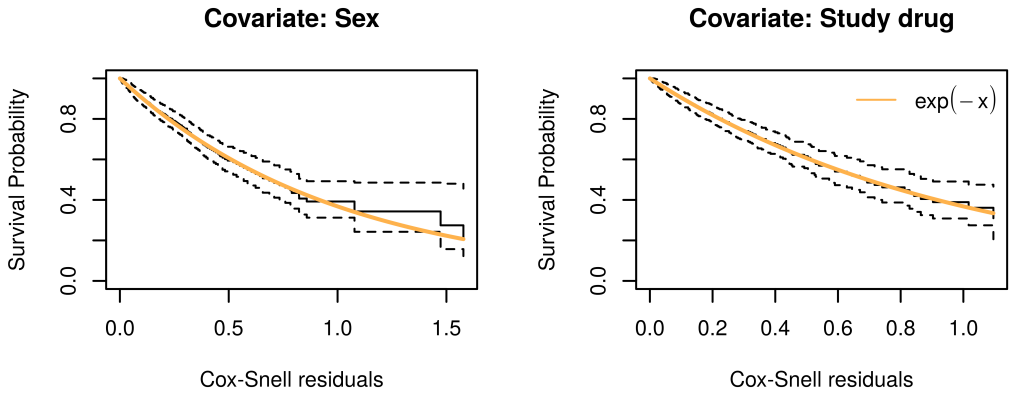


Figure 6.2: Cox-Snell residuals for two separate applications to the PBC data. The solid black line is the Kaplan Meier estimate for the survival function, and dashed black lines its 95% confidence interval. The overlaid orange line is the survival function of the unit exponential distribution.

6.2 Hypothesis testing and model selection for joint models

Joint models enjoy readily-available inference: The model fit under maximum likelihood estimation carried out as described in Section 2.2 leads directly to standard likelihood inference tests. For instance, we can employ the usual Wald statistic $W = \hat{\beta}_x/\mathrm{SE}(\hat{\beta}_x)$ for some generic fixed-effect $\hat{\beta}_x \in \hat{\boldsymbol{\beta}}$, testing for significance of this parameter by vetting of its corresponding p-value obtained from the standard normal distribution

$$p_W = 2\Pr(|W| > z)\,,$$

where the above described procedure is analogous to the comparison of $\hat{\beta}_x^2/\mathrm{Var}[\hat{\beta}_x]$ to the χ^2 distribution.

We can compare two *nested* fitted joint models $\mathcal{M}_0$ and $\mathcal{M}_1$ – the former nested within the latter – by means of the likelihood ratio test (LRT),

$$\begin{aligned} \mathrm{LRT} &= -2\log(\ell(\mathcal{M}_1) - \ell(\mathcal{M}_0)) \\ \mathrm{LRT} &\sim \chi^2_{P(\mathcal{M}_1)-P(\mathcal{M}_0)} \end{aligned} \tag{6.4}$$

where $\ell(\cdot)$ is the log-likelihood evaluated at $\{\hat{\boldsymbol{\Omega}}, \hat{\boldsymbol{b}}_i\}$ from each of the fitted models and $P(\cdot)$ returns the number of parameters i.e. the length of the parameter vector (2.30). In addition, the LRT could be used in an analogous manner on a 'per-parameter' basis to the Wald test previously described, at (greatly) increased computational expense. Hereafter we compare two nested joint models by LRT and appraise individual parameter significance by the Wald test.

One drawback for the comparison of two joint models – moreover two GLMMs – fit under different random effects specifications is that critical issues arise when testing whether or not the additional random effect(s) should be included. Inclusion of more random effects will account for (extra) variability amongst the subjects, only serving to increase the dimension of D. The LRT carried out on the two models then essentially sets an element of diag(D) (i.e. the variance) to zero, thereby precluding standard asymptotic inference since this value exists outside of its parameter space (Rizopoulos, 2012b).

In the nested case (e.g. when *only* additional fixed effects, or survival covariates distinguish the two joint models), then the LRT laid-out in (6.4) is sufficient in determining the superior model. If the two models are *not* nested, due to presence of additional longitudinal processes and/or more complex random effects structures, these standard asymptotic tests are inappropriate, and instead standard information criteria are used to compare these non-nested joint models. Namely, we consider the ubiquitous Akaike's information criterion (AIC, Akaike (1974)) and Bayesian information criterion (BIC, Schwarz (1978)), defined for a fitted joint model $\mathcal{M}$ as

$$\begin{aligned} \text{AIC} &= -2\ell(\mathcal{M}) + 2P(\mathcal{M}) \\ \text{BIC} &= -2\ell(\mathcal{M}) + P(\mathcal{M}) \log n_{\text{obs}}, \end{aligned} \tag{6.5}$$

with $\ell(\cdot)$ and $P(\cdot)$ defined earlier and n_{obs} the number of observations in the data (i.e. the number of 'rows'). Lower values are preferable across both criteria. These criteria usually 'prefer' (or 'declare best') different models: AIC penalises model complexity far less harshly than BIC, i.e. generally speaking the penalty term under BIC is larger than AIC's, $P(\mathcal{M}) \log n_{\text{obs}} > 2P(\mathcal{M})$, leading to BIC preferring simpler models in general; especially when the sample size is relatively small.

In reality then, selecting between two competing joint models involves careful consideration of a multitude of factors, including the information criteria (6.5); behaviour of the standardised residuals for the longitudinal process(es) as introduced in Section 6.1.1, as well as checking conformity of the Cox-Snell residuals discussed in Section 6.1.2. The user also needs to contemplate whether or not the two candidate joint models are indeed nested or not.

In the next section, we introduce prognosis based upon a fitted joint model, and in Section 6.4 introduce measures to quantify predictive accuracy; which may be additional characteristics we contemplate in ascertaining the preferred joint model.

6.3 Dynamic predictions

Thus far, we have largely considered joint models in vacuo: Focussing on parameter estimates, particularly for the association parameters $\hat{\boldsymbol{\gamma}}$, with allusions of basic statistical inference upon them. One avenue for using a joint model (or indeed any statistical model) is to obtain predictions for some outcome of interest: In our case, survival.

Rizopoulos (2011) outlined one such method for bridging between a fitted joint model and a prognostic one, producing predictions of survival probabilities. Owing to the longitudinal nature of the data, potentially important information may be added a 'future' follow-up time; the *dynamic* nature of the predictions are a unique selling point over e.g. simply using a Cox model.

These so-called '*dynamic predictions*' allow us to estimate the probability that a certain subject i, with longitudinal information available up to follow-up time t, survives future time u. This subject-specific prediction is then particularly appealing in an era of personalised medicine (Rizopoulos, 2012b).

More formally, consider having successfully fitted joint model to the 'full' set of observed data $\mathcal{D} = \{\mathcal{D}_i,\ i = 1, \ldots, n\}$, with resultant parameter estimates $\hat{\boldsymbol{\Omega}}$. Then, we want to estimate for a *new* subject i with the generic set of observed data up to time t, $\mathcal{D}_i(t) = \{\boldsymbol{Y}_{i1}(t), \ldots, \boldsymbol{Y}_{iK}(t), \boldsymbol{S}_i\}$ with all other design measures (e.g. $\mathrm{X}_{ik}(t)$) holding implicit membership,

$$\pi_i(u|t) = \Pr\left(T_i^* \geq u | T_i^* > t, \mathcal{D}_i(t), \mathcal{D}; \hat{\boldsymbol{\Omega}}\right). \tag{6.6}$$

The *dynamic* property materialises owing to the conditioning on follow-up time t, since $\pi_i(u|t) \neq \pi_i(u|t+\delta)$ for some period of follow-up δ having occurred (i.e. new information becomes available).

6.3.1 Estimation of $\pi_i(u|t)$

For a given follow-up time t we have the previously defined set of observed data up to time t, $\mathcal{D}_i(t)$. Evaluation of the probability of interest (6.6) on the *observed data* requires we utilise the conditional independence of the longitudinal process(es) and the survival

outcome given the random effects (2.3) to obtain (Rizopoulos, 2012b)

$$\begin{aligned}\pi_i(u|t) &= \Pr\Big(T_i^* \geq u | T_i^* > t, \mathcal{D}_i(t)\,; \hat{\boldsymbol{\Omega}}\Big)\,, \\ &= \int \Pr\Big(T_i^* \geq u | T_i^* > t, \mathcal{D}_i(t)\,, \boldsymbol{b}_i; \hat{\boldsymbol{\Omega}}\Big)\, f\Big(\boldsymbol{b}_i | T_i^* > t, \mathcal{D}_i(t)\,; \hat{\boldsymbol{\Omega}}\Big)\, d\boldsymbol{b}_i, \\ &= \int \frac{S(u|\boldsymbol{b}_i, \mathcal{D}_i(u)\,; \hat{\boldsymbol{\Omega}})}{S(t|\boldsymbol{b}_i, \mathcal{D}_i(t)\,; \hat{\boldsymbol{\Omega}})} f\Big(\boldsymbol{b}_i | T_i^* > t, \mathcal{D}_i(t)\,; \hat{\boldsymbol{\Omega}}\Big)\, d\boldsymbol{b}_i, \end{aligned} \tag{6.7}$$

with the survival function $S(t|\cdot)$ denoting the probability of survival past time t given by

$$S(t|\boldsymbol{b}_i, \mathcal{D}_i(t)\,; \hat{\boldsymbol{\Omega}}) = \exp\left[- \int_0^t \lambda_0(t) \exp\left\{ \boldsymbol{S}_i^\top \boldsymbol{\zeta} + \sum_{k=1}^{K} \gamma_k \boldsymbol{W}_k(t)\, \boldsymbol{b}_{ik} \right\} dt \right]. \tag{6.8}$$

We can appraise the quantity (6.7) using both a first-order estimate as well as Monte Carlo simulation, which we explore in upcoming sections.

6.3.2 First-order estimate for $\pi_i(u|t)$

We can employ the empirical Bayes estimator for the data available up to time t, defined as

$$\hat{\boldsymbol{b}}_i^{(t)} = \underset{\boldsymbol{b}_i}{\operatorname{argmax}} \Big\{ \log f\Big(\boldsymbol{b}_i, \mathcal{D}_i(t)\,|T_i^* > t; \hat{\boldsymbol{\Omega}}\Big) \Big\}\,, \tag{6.9}$$

i.e. in a similar spirit to (3.2) with additional conditioning on the available data which forms the complete data log-likelihood in (6.9).

Previously Rizopoulos (2011) found the modal estimate $\hat{\boldsymbol{b}}_i^{(t)}$ utilising available longitudinal information *only*. However, in keeping with methodologies employed in the thesis outlined (3.2), and as carried out by existing software packages (`joineRML`, Hickey et al. (2018a)), we include the survival sub-model's contribution to this complete data log-likelihood in calculation of $\hat{\boldsymbol{b}}_i^{(t)}$.

Substituting $\hat{\boldsymbol{b}}_i^{(t)}$ into (6.7) we obtain the approximated probability

$$\tilde{\pi}_i(u|t) = \frac{S(u|\hat{\boldsymbol{b}}_i^{(t)}, \mathcal{D}_i(u); \hat{\boldsymbol{\Omega}})}{S(t|\hat{\boldsymbol{b}}_i^{(t)}, \mathcal{D}_i(t); \hat{\boldsymbol{\Omega}})} \tag{6.10}$$

The estimates produced by (6.10) perform well in practice (Rizopoulos, 2011). However, establishing a handle on variability of $\tilde{\pi}_i(u|t)$, is difficult since the variability is effectively 'double-barrelled', i.e. one must account for both $\text{Var}[\hat{\boldsymbol{\Omega}}]$ and $\text{Var}[\hat{\boldsymbol{b}}_i^{(t)}]$. If uncertainty in the approximated point estimate(s) of $\tilde{\pi}_i(u|t)$ are of interest (which they most likely are),

then it is perhaps wiser to undertake the Monte Carlo scheme outlined in the next section.

6.3.3 Estimate for $\pi_i(u|t)$ by Monte Carlo simulation

Recalling the quantity (6.6), we note Rizopoulos (2011) considered its expected value taken against the posterior probability density of the parameters given the data to which the joint model was fit,

$$\begin{aligned}\pi_i(u|t) &= \Pr(T_i^* \geq u | T_i^* > t, \mathcal{D}_i(t)\,, \mathcal{D}; \boldsymbol{\Omega})\,, \\ &= \int \Pr(T_i^* \geq u | T_i^* > t, \mathcal{D}_i(t)\,; \boldsymbol{\Omega}) f(\boldsymbol{\Omega}|\mathcal{D}) d\boldsymbol{\Omega}, \end{aligned} \tag{6.11}$$

i.e. its posterior expectation.

In the above, the quantity $\Pr(T_i^* \geq u | T_i^* > t, \mathcal{D}_i(t)\,; \boldsymbol{\Omega})$ is given by the ratio of survival functions averaged over the random effects distribution previously written (6.7). Attention then sensibly turns to a candidate distribution which well-approximates $f(\boldsymbol{\Omega}|\mathcal{D})$. For a suitably large sample size n, Rizopoulos (2011) utilise the multivariate normal centered at $\hat{\boldsymbol{\Omega}}$ with variance-covariance $\mathcal{I}^{-1}\big(\hat{\boldsymbol{\Omega}}\big)$, with $\mathcal{I}$ being approximated by methods described in Section 2.4.

In order to obtain $l = 1, \ldots, N$ realizations of the conditional survival probability $\pi_i(u|t)$, and compute summaries of interest, we proceed with the following Monte Carlo scheme:

1. Draw $\boldsymbol{\Omega}^{(l)} \sim N\big(\hat{\boldsymbol{\Omega}}, \mathcal{I}^{-1}\big(\hat{\boldsymbol{\Omega}}\big)\big)$;
2. Draw $\hat{\boldsymbol{b}}_i^{(l)} \sim \boldsymbol{b}_i | T_i^* > t, \mathcal{D}_i(t)\,; \boldsymbol{\Omega}^{(l)}$;
3. Calculate the l^{th} ratio of conditional survival probabilities:

$$\pi_i^{(l)}(u|t) = \frac{S\big(u|\hat{\boldsymbol{b}}_i^{(l)}, \mathcal{D}_i(u); \boldsymbol{\Omega}^{(l)}\big)}{S\big(t|\hat{\boldsymbol{b}}_i^{(l)}, \mathcal{D}_i(t); \boldsymbol{\Omega}^{(l)}\big)};$$

4. Repeat steps 1.–3. N times and compute relevant summary statistics from the obtained probabilities $\boldsymbol{\pi}_i(u|t) = \Big(\pi_i^{(1)}(u|t)\,, \ldots, \pi_i^{(N)}(u|t)\Big)^\top$.

Sampling from the multivariate normal $N\big(\hat{\boldsymbol{\Omega}}, \mathcal{I}^{-1}\big(\hat{\boldsymbol{\Omega}}\big)\big)$ is trivial, and the draw for $\hat{\boldsymbol{b}}_i$ is recommended to be carried out from the multivariate t distribution, $t_4\big(\hat{\boldsymbol{b}}_i, \hat{\Sigma}_i\big)$, with location $\hat{\boldsymbol{b}}_i$ previously given (6.9) and variance $\hat{\Sigma}_i$ found by e.g. (3.3). In practise the random effects sampling is carried out by a Metropolis-Hastings scheme.

6.4 Prognostic accuracy measures for joint models

With routines to obtain survival probabilities established, we now seek to lay out measures of predictive performance which utilise and appraise said probabilities. Such performance measures are invaluable tools to e.g. clinical practitioners, assigning levels of confidence when informing interested parties – facilitating medical decision making in an effort to improve health outcomes – of clinical predictions (van Smeden et al., 2021). We briefly declare our interest in *prognostic* prediction, rather than *diagnostic* prediction. van Smeden et al. (2021) neatly distinguish the latter here as being cross-sectional in nature, whereas the former is longitudinal; marrying-up with the *dynamic* predictions we introduced in the previous section.

The quality of a fitted joint model – that is, both the longitudinal and survival parts – can be inferred from information criteria previously outlined in (6.5); allowing for comparison of alternate models fit to the same data. However, interest may fall on the event-time, and within the joint modelling context this provides insight into how well the longitudinal measure(s) can predict this survival outcome.

Rizopoulos (2012b) report something of a schism between calibration and discrimination measures. The former assesses how well the predicted probabilities from the model match the actual outcomes; i.e. evaluating whether the predicted probabilities are accurate estimates of the true likelihood of event occurrence. For example, Henderson et al. (2002) utilise a calibration-based approach to verify three candidate models fit to cirrhosis data, reporting the agreement between the observed Kaplan-Meier plots and those generated by simulation from each fitted joint model. On the other hand, discrimination measures evaluate how well the fitted model can distinguish between those groups of subjects who experience the event and those who do not. These discrimination approaches give rise to e.g. the widely-digested receiver operating characteristic (ROC) curves, which we go on to detail in Section 6.4.2; calibration measures are outlined separately in Section 6.4.3.

6.4.1 Setting out follow-up windows and probabilities of interest

Consider we have subject(s) each having longitudinal information available up to predetermined follow-up time T_{start}, and denote these subjects $i = 1, \ldots, n_{\text{alive}}$, $n_{\text{alive}} = \sum_{i=1}^{n} I(T_i^* > T_{\text{start}})$. We wish to investigate whether each of the n_{alive} subjects survive past a future *horizon* time-point $h = T_{\text{start}} + \delta$, with δ some period of elapsed time. The idea being that a practitioner could intervene for certain subjects if their survival outcome was deemed poor. We can then determine discriminatory capabilities of the fitted joint model in the window of follow-up $w = (T_{\text{start}}, h]$.

The time-window of interest w provides an f-vector of failure times, $\boldsymbol{u}_w = (u_1, \ldots, u_f)^\top, T_{\text{start}} < u_j \leq h \; \forall \; j = 1, \ldots, f$, which occur in the set of data the joint model was fit to $\mathcal{D}$ between the start and horizon times. The routines carried out in Sections 6.3.2 and 6.3.3 therefore produce f conditional probabilities for each of the n_{alive} subjects: $\hat{\pi}_i(u_1|T_{\text{start}}), \ldots, \hat{\pi}_i(u_f|T_{\text{start}})$. However, since we are interested only in failure *within* the window w (or equivalently, survival *past* future time-point h), we elect some summary measure on these probabilities, say $g(\cdot)$, and produce *one* probability for the window w for *each* subject

$$\hat{\pi}_i(w) = g(\hat{\pi}_i(u_1|T_{\text{start}}), \ldots, \hat{\pi}_i(u_f|T_{\text{start}})).$$

In practise one may set $g(\cdot)$ to return the minimum value i.e. the 'worst-case' scenario for subject i; or to simply return the probability calculated for the final candidate follow-up time i.e. probability of surviving the window. If the latter is chosen here then we note $\hat{\pi}_i(w) = \hat{\pi}_i(u_f|T_{\text{start}})$.

Next, we can define the usual 'hit'/error outcomes in what are effectively binary terms. We introduce candidate thresholding probabilities $\boldsymbol{c} = (0.00, 0.01, 0.02, \ldots, 0.99, 1.00)^\top$ against which $\hat{\pi}_i(w)$ can be compared to declare subject i having failed/survived in w at different probabilistic thresholds $c_j \in \boldsymbol{c}$. Namely, we begin with quantifying the correct identification of a failure in w (true positive) and correct declaration of survival past u_f (true negative), as determined by $\hat{\pi}_i(w)$ and c_j (Andrinopoulou et al., 2021)

$$\begin{aligned} \text{True Positive} &= \Pr(\hat{\pi}_i(w) \leq c_j | T_{\text{start}} < T_i^* < h), \\ \text{True Negative} &= \Pr(\hat{\pi}_i(w) > c_j | T_i^* > u_f). \end{aligned} \tag{6.12}$$

In reality, the example measures in (6.12) as well as further measures consolidated in the next section are found column-wise using the $n_{\text{alive}} \times 101$ matrix

$$\mathrm{M} = \hat{\pi}_i(w) \overset{\leq}{\otimes} \boldsymbol{c},$$

where $\boldsymbol{x} \overset{\leq}{\otimes} \boldsymbol{y}$ denotes the outer product of the elements of vectors $x_i \in \boldsymbol{x}$ and $y_j \in \boldsymbol{y}$ with the $(i,j)^{\text{th}}$ element of the resultant matrix taking value 1 if $x_i \leq y_j$ and zero otherwise. Moreover, the rows of M contain $\hat{\pi}_i(w), i = 1, \ldots, n_{\text{alive}}$ compared against threshold probability $c_j \in \boldsymbol{c}$, which construct its columns.

Remark. One notes by the definition of w that these classifiers, furthermore their ensuing contributions to ROC and AUC measures in the next section, are time-dependent in nature. This means that the capability of a biomarker (or indeed, the 'whole' fitted joint model) to distinguish between those who will/won't fail over a certain period of follow-up may well improve – or indeed deteriorate – as follow-up progresses. Under the joint

modelling approach then (compared to e.g. diagnostic prediction by logistic regression, or prognostic prediction via a Cox model using only baseline covariates), the evolution of the exogenous biomarker(s) is an important feature of follow-up.

6.4.2 Discrimination measures

We first introduce the 'true failure' indicator function $\Delta_i(w) = I(T_{\text{start}} < T_i \leq h) \cap I(\Delta_i = 1)$, with w defined in the previous section, and the 'predicted failure' indicator function determined by some probabilistic threshold c_j, $\hat{\Delta}_i(w, c_j) = I(\hat{\pi}_i(w) \leq c_j)$. We 'build on' the true positives and negatives described in (6.12) and complete the usual contingency table presented in Table 6.1. We briefly note that false negatives constitute type II error, and false positives type I.

		Predicted status $\hat{\Delta}_i(\cdot)$	
		Positive	Negative
Actual $\Delta_i(\cdot)$	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Table 6.1: 2×2 contingency table for outcomes of the (pseudo-)binary classifier of survival in time-window w. 'Positive' refers to the indicator function(s) $\Delta_i(\cdot)$ and $\hat{\Delta}_i(\cdot)$ returning value 1 i.e. failure (is deemed to have) occurred.

A swathe of further evaluation metrics relating to the probability of detection, and accuracy, are derived purely based on the four outcomes in Table 6.1 and are presented in Table 6.2. We additionally consider two proffered summary measures, which aim to capture something of a 'trade-off' amongst the measures given in Table 6.2. The first, Youden's J statistic, J_Y, aims to provide a single statistic for a dichotomous (i.e. failure/survival) outcome (Youden, 1950), with larger values better. Secondly, the F_1 score is another measure of the test's accuracy:

$$\begin{aligned} J_Y &= \text{Sensitivity} + \text{Specificity} - 1 = \frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}} - 1 \\ F_1 &= \frac{2\text{PPV} \times \text{Sensitivity}}{\text{PPV} + \text{Sensitivity}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}; \end{aligned} \tag{6.13}$$

since we consider many probabilistic thresholds to obtain measures of discriminatory performance, these two summary measures could be used to identify the best-performing candidate probability threshold $c_j \in \boldsymbol{c}$. We note that the F_1 score has come under some criticism in wider literature, one such example being Hand and Christen (2018), and this identification of the 'best' probability threshold will be carried out using J_Y.

The ROC curve provides a visual representation of the true positive rate (the proportion of correctly predicted failures out of all actual occurring failures), and the false positive

Name (abbv.)	Formula	Description
True positive rate (TPR)	$\frac{\text{TP}}{\text{TP+FN}}$	**Sensitivity**: Probability that a true positive will test positive i.e. $\Delta_i(\cdot) = 1 \cap \hat{\Delta}(\cdot) = 1$
False positive rate (FPR)	$\frac{\text{FP}}{\text{FP+TN}}$	**1-Specificity**: Probability that a true negative will test positive i.e. $\Delta_i(\cdot) = 0 \cap \hat{\Delta}_i(\cdot) = 1$
Accuracy (Acc.)	$\frac{\text{TP+TN}}{n_{\text{alive}}}$	Proportion of correct predictions (both true positives and true negatives) among the total number of cases examined
Positive predictive value (PPV)	$\frac{\text{TP}}{\text{TP+FP}}$	Probability that those deemed to experience the event experience it in actuality. Sometimes called **precision**
Negative predictive value (NPV)	$\frac{\text{TN}}{\text{TN+FN}}$	Probability that those deemed to *not* experience the event do not experience it in actuality.

Table 6.2: Further evaluation metrics to be calculated using the four outcomes in Table 6.1.

rate (the proportion of false positives out of all actual non-failures). The curve itself is generated by the probability 'grid' $\boldsymbol{c}$, which results in adjustments to the FPR and TPR. A ROC curve from a 'dummy' example fit to the PBC data, with sole covariate of drug recipiency is given in Figure 6.3. The area under the ROC curve ('AUC-ROC', or simply 'AUC') is determined by the integral

$$\text{AUC} = \int_0^1 \text{ROC}(c)\, dc, \tag{6.14}$$

and provides a summary measure of the predictive accuracy index of the model. In practise we calculate the AUC using the formula for the area under a trapezoid (half of the length of the parallel sides (FPR) multiplied by the height (TPR)) calculated at each consecutive probability threshold 'pair' c_j and c_{j+1}

$$\text{AUC} = \sum_{j=1}^{100} \frac{1}{2} \times (\text{TPR}_j + \text{TPR}_{j+1}) \times (\text{FPR}_{j+1} - \text{FPR}_j)\,. \tag{6.15}$$

We expect AUC to lie between 0.5 and 1.0, with higher values indicating better discriminatory power; an AUC less than 0.5 signifies that flipping a coin would provide better discrimination between subjects. Lastly, the dashed line in Figure 6.3 – representing J_Y – indicates for this example model that the optimal trade-off between sensitivity (TPR = 0.512) and specificity (FPR = 0.091) occurs at $c_j = 0.21$; a balance is struck between correctly identifying positive and negative cases. That is, if $\pi_i(u|t) \leq 0.21$ then

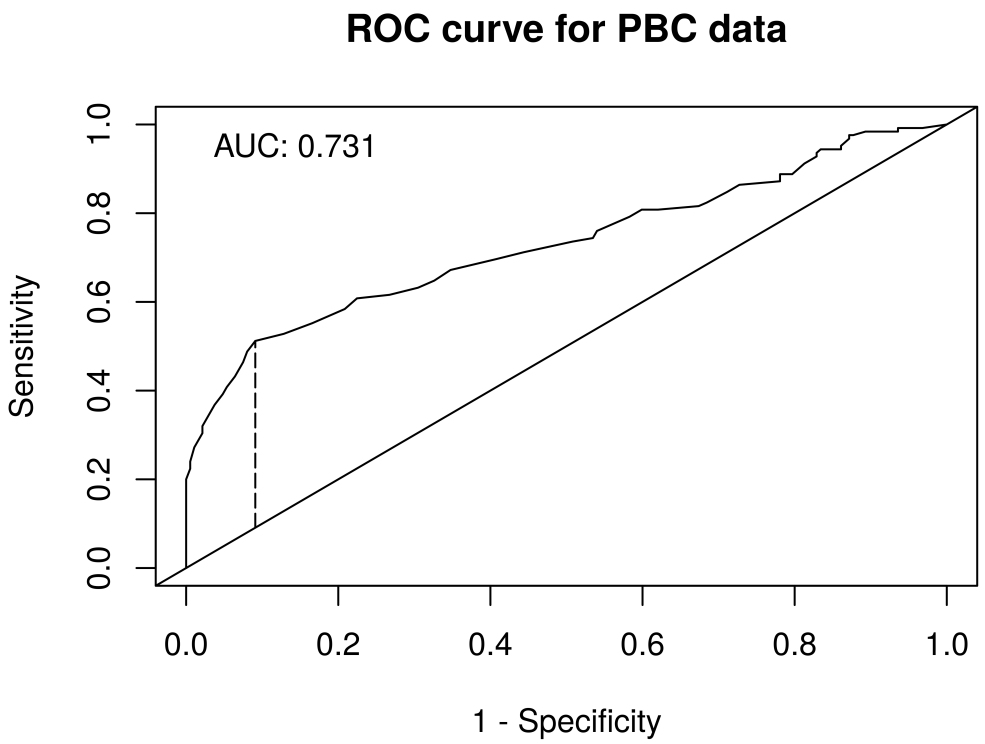


Figure 6.3: Example receiver operator characteristic (ROC) curve obtained from a Cox PH model fit to PBC data. The dashed line indicates the minimal Y_J value and was obtained for $c_j = 0.210$

we can be reasonably confident that there are grounds for intervention, given the overall *decent* discriminatory ability of the model (AUC = 0.731), with the aforementioned 'optimal' trade-off between sensitivity and specificity for this example observed here.

6.4.3 Calibration measures

Finally, we outline calibration measures alluded to in Section 6.4. Calibration for survival outcomes is based on the expected error of forecasting future events. As was remarked upon in the Section 6.4.1, these measures are time-dependent in nature since they take into account the dynamic nature of the longitudinal outcome.

Setting out our calibration metrics in a similar 'windows' as our discriminatory ones, we define the prediction error in a similar manner to e.g. Rizopoulos et al. (2017)

$$\text{PE}(w) = \mathbb{E}[\mathcal{L}(I(T_i^* > h) - \hat{\pi}_i(w))]\,, \tag{6.16}$$

where $\mathcal{L}$ is a loss function of choice. We note that the predictive error is the same as

the Brier score when the square loss function is chosen. We utilise the estimate given by Henderson et al. (2002) for $\text{PE}(w)$ which accounts for censoring,

$$\begin{aligned}\widehat{\text{PE}}(w) = \frac{1}{n_{\text{alive}}} \sum_{i=1}^{n_{\text{alive}}} & I(T_i > h)\mathcal{L}(1 - \hat{\pi}_i(w)) + \Delta_i(w)\,\mathcal{L}(0 - \hat{\pi}_i(w)) \\ & + C_i(w)(\hat{\pi}_i(w)\,\mathcal{L}(1 - \hat{\pi}_i(w)) + (1 - \hat{\pi}_i(w))\,\mathcal{L}(0 - \hat{\pi}_i(w)))\end{aligned} \tag{6.17}$$

where $C_i(w) = I(T_{\text{start}} < T_i \leq h) \cap I(\Delta_i = 0)$ i.e. censored in the window of interest. The first term measures how well the model predicts survival, contributing to the prediction error when the subject survives past horizon time h with the loss function $\mathcal{L}$ quantifying accuracy of the survival prediction. The second term similarly contributes how well the model predicts failures in the window to the prediction error. The final term adjusts the prediction error for those who were censored; needing to account for both the possibility they survived or failed. Within this censoring adjustment the quantity $\hat{\pi}_i(w)\,\mathcal{L}(1 - \hat{\pi}_i(w))$ assesses the model's prediction error when the subject is censored but predicted to survive, with the remaining term its complement. In summary, (6.17) aggregates the prediction error over those who survived past time h, failed in the window w, or were censored in the window.

Rizopoulos et al. (2017) point out that this estimated measure (6.17) can be used to compare two nested joint models $\mathcal{M}_0$ and $\mathcal{M}_1$ – the former nested within the latter – by

$$R(w) = 1 - \frac{\widehat{\text{PE}}_{\mathcal{M}_1}(w)}{\widehat{\text{PE}}_{\mathcal{M}_0}(w)}, \tag{6.18}$$

this quantity R measures how much the additional information (e.g. an additional biomarker) used to fit $\mathcal{M}_1$ improves prognostic accuracy by in the window w.

6.4.4 Correcting for optimism

Unfortunately, simply evaluating the prognostic performance for the fitted joint model $\mathcal{M}$ on the same data to which the model was fit, $\mathcal{D}$, will lead to estimates for e.g. $\text{AUC}(\cdot)$ which are optimistic in nature (Andrinopoulou et al., 2021; van Smeden et al., 2021): Essentially, the model $\mathcal{M}$ has adequately captured e.g. trajectories and relationships in the data to which it is evaluated against, leading to these inflated estimates.

We correct for this optimism via internal validation, specifically by a bootstrapping approach given in Andrinopoulou et al. (2021). For a pre-determined window of interest w, we have the following quantities which are calculated from the joint model fit $\mathcal{M}$ on the original data, $\text{AUC}_{\mathcal{M}}(w)$ and $\widehat{\text{PE}}_{\mathcal{M}}(w)$. We then seek to create $b = 1, \ldots, B$ corrected measures by the following scheme:

1. Create a bootstrap sample of the data, $\mathcal{D}^{(b)}$, and (re)fit the joint model, obtaining $\mathcal{M}^{(b)}$.

2. Calculate the prognostic performance measures using $\mathcal{M}^{(b)}$ on $\mathcal{D}^{(b)}$, obtaining $\text{AUC}_b(w)$ and $\widehat{\text{PE}}_b(w)$.

3. Calculate the prognostic performance measures using $\mathcal{M}^{(b)}$ on $\mathcal{D}$, obtaining $\text{AUC}_{b^*}(w)$ and $\widehat{\text{PE}}_{b^*}(w)$.

4. Calculate the optimism o in the b^{th} replicate

$$\begin{aligned} \text{AUC}_o(w) &= \text{AUC}_b(w) - \text{AUC}_{b^*}(w)\,, \\ \widehat{\text{PE}}_o(w) &= \widehat{\text{PE}}_b(w) - \widehat{\text{PE}}_{b^*}(w)\,. \end{aligned}$$

5. Calculate the b^{th} corrected estimate for the prognostic performance measures

$$\begin{aligned} \text{AUC}^{(b)}_{\text{corr}}(w) &= \text{AUC}_{\mathcal{M}}(w) - \text{AUC}_o(w)\,, \\ \widehat{\text{PE}}^{(b)}_{\text{corr}}(w) &= \widehat{\text{PE}}_{\mathcal{M}}(w) + \widehat{\text{PE}}_o(w)\,. \end{aligned}$$

Carrying out the above yields B corrected estimates of our two considered measures, upon which we can form summary measures and investigate e.g. boxplots of estimates. In order to circumvent some of the increased computational cost incurred by fitting B bootstrapped models, we commence their EM algorithms at the MLEs $\hat{\boldsymbol{\Omega}}$ from $\mathcal{M}$, and set lower convergence criteria thresholds $\xi_1 = 5 \times 10^{-3}$ and $\xi_2 = 1 \times 10^{-2}$ in (3.4). Additionally, convergence is declared when *either* the absolute or relative criterion are satisfied.

When fitting the b^{th} model $\mathcal{M}^{(b)}$ back on the original data $\mathcal{D}$, it's possible that some subject i is *not* sampled thereby absent in $\mathcal{D}^{(b)}$, potentially omitting an observed failure time in the process. In these scenarios, the 'missing' random effects are calculated by (3.2) with variance (3.3), wherein $\boldsymbol{\Omega}$ is substituted by $\hat{\boldsymbol{\Omega}}^{(b)}$. The baseline hazard for the original data given $\mathcal{M}^{(b)}$ is then calculated by (3.13). The above routine can additionally be used to find the improvement in prognostic accuracy (6.18), where the bootstrapped data $\mathcal{D}^{(b)}$ is used to fit *both* competing models, predictive error calculated on both bootstrapped and original models, before the original $R(w)$ is corrected for optimism B times.

Chapter 7

Application: Primary Biliary Cirrhosis

7.1 Introduction and motivation

We now bring together the (flexible) joint models we established in Chapters 2–4 with post-hoc analyses such as model selection and dynamic prediction we described in Chapter 6 in an exemplary application to the Mayo Clinic primary billiary cirrhosis (PBC) clinical trial (Murtaugh et al., 1994) first mentioned in Section 1.3. Conducted over ten years, the PBC trial data is a popular example in joint modelling literature owing to the presence of many longitudinal biomarkers – moreover, it is particularly appealing to us due to the mixture of response types – and information on an event-time (death or transplantation). Additionally, there is an opportunity to 'verify' results indicating a biomarkers' association with mortality (Murtaugh et al., 1994) which pre-date joint modelling.

The intention for this chapter is to act as a 'full' application to this oft-used clinical data set in which we demonstrate how one may arrive at the 'best-fitting' joint model for the data using the methodology outlined thus far. We begin by providing a detailed overview of the data itself in Section 7.2, before turning our attention to model building in Section 7.3. We initiate this by first finding the best-fitting survival sub-model (i.e. by Cox PH alone); performing then a similar process for the longitudinal biomarkers we consider. Next, in Section 7.4 we use these best-fitting longitudinal and survival sub-models in a series of univariate joint model fits to test for standalone significant associations with mortality; followed by the multivariate scenario which supplements our arrival at the most parsimonious joint model in Section 7.5. Finally, in Section 7.6 we establish the prognostic capabilities of our chosen model using methods from Section 6.4. The `R` code used in this Chapter is available at `https://github.com/jamesmurray7/thesis/tree/`

`main/thesis-app/PBC`

Remark. These first steps of the model building process occur outside of the joint modelling framework to mimic what practitioners may do in practise: Elucidating the 'best' available fit for each of the longitudinal and survival sub-models, before jointly analysing these in a (multivariate) joint model.

7.2 Data description and exploration

There are $n = 312$ subjects in the PBC data, of whom $n = 154$ (49.4%) were randomised to placebo treatment and the remainder to the active treatment D-penicillamine. Patients were monitored until either they experienced mortality, underwent liver transplantation, or reached the end of follow-up. We are interested with the clinical outcome of mortality *only*. In total, $n = 140$ (44.9%) subjects died during follow-up. The usual Kaplan-Meier curve for the data is presented in Figure 7.1 along with a histogram of failure times. We observe from said histogram that many subjects die in the study's infancy.

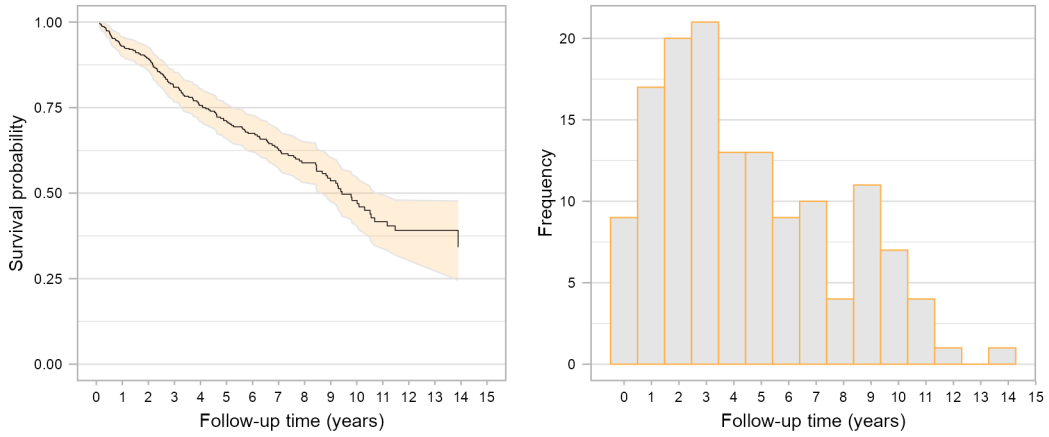


Figure 7.1: Left-hand plot: Kaplan-Meier estimate of the survival function for the PBC data. The shaded ribbon signifies the 95% confidence interval for the survival function. Right-hand plot: Histogram of *failure times* occurring over follow-up.

As mentioned, numerous longitudinal outcomes exist with varying degrees of completeness in the data. Of these, we consider four to be continuous: Log serum bilirubin; log serum aspartate aminotransferase ('AST'); serum albumin and prothrombin time. Three are binary markers which indicate presence of: Enlarged liver (hepatomegaly); accumulation of fluid in abdomen (ascites) and malformed blood vessels in skin ('spiders'). Finally, platelet count and alkaline phosphatase are treated as count biomarkers. The longitudinal trajectories of these non-binary biomarkers is given in Figure 7.2. Here we observe separation between average trajectories amongst those who did/not survive follow-up, perhaps most

notably for (log) serum bilirubin and albumin.

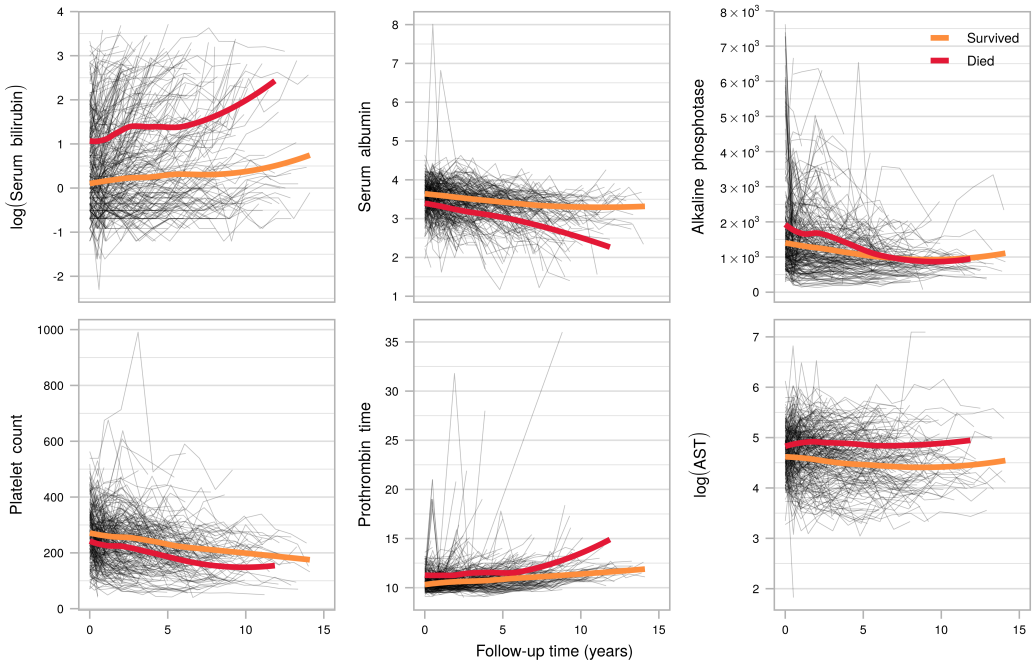


Figure 7.2: Longitudinal trajectories for continuous or count biomarkers present in the PBC data. Grey lines show individual trajectories and overlaid smoothed (LOESS) curves the average trajectories for those who experienced mortality during follow-up and those who did not.

The binary biomarkers are presented in Appendix C.3.1. Here we note the relative lack-of prevalence for spiders and ascites, especially compared with hepatomegaly. With this in mind, especially in conjunction with our conclusions in Chapters 4 and 5 – that the approximation may not suit, or struggle to fit, a binary response – we elect to *only* consider hepatomegaly of these three binary biomarkers hereafter.

We have access to many baseline covariates which could be of interest to both the survival and longitudinal processes. We consider the (standardised) age at baseline; the patient's sex; recipiency of drug and the histologic status of the disease. The histological stage is an ordinal covariate with four stages reflecting worsening disease progression, from stage 1 indicating presence of lesions in the bile duct to stage 4 cirrhosis of the liver (Purohit and Cappell, 2015).

A data characteristics table split by those who survived the follow-up period is given in Table 7.1. Here we note from a precursory glance at the data that those who died tended to be older and enter the study at a more advanced cirrhosis disease state. The analysis set used here and in all subsequent analyses are those with *no* missing covariate or biomarker values.

Remark. The presence of 'cirrhosis' as an histologic state given our naming of 'primary

billiary cirrhosis' is perhaps confusing. Indeed 'cirrhosis' is a feature *only* of advanced disease, such that recently patient advocacy groups have proposed a changing of its name to 'primary billiary cholangitis' to more accurately reflect the disease's pathology (Beuers et al., 2015). In keeping with the vast majority of existing literature in joint modelling, we solely note this contention in naming and proceed as before.

	Transplanted or survived ($n = 172$)	Died ($n = 140$)	p-value
n (%) Female	162 (94.2%)	114 (81.4%)	0.001
n (%) received placebo	85 (49.4%)	69 (49.3%)	0.999
Median [IQR] follow-up length	6.5 [4, 10]	4 [2, 7]	< 0.001
Mean (SD) age	47.07 (9.876)	53.64 (10.323)	< 0.001
n (%) histologic status			< 0.001
1	15 (8.7%)	1 (0.7%)	
2	46 (26.7%)	21 (15.0%)	
3	72 (41.9%)	48 (34.3%)	
4	39 (22.7%)	70 (50.0%)	

Table 7.1: Baseline data characteristics for the PBC data, stratified by whether or not the subject died during follow-up, or received a transplant or reached the end of the follow-up period. The p-values are calculated from appropriate tests for difference between clinical endpoint. Where the characteristic of interest reported as median [IQR], Wilcoxon rank-sum test was used; where mean (SD), student t-test was used; and where categorical, χ^2 test was used.

7.3 Model building

With the data summarised in the previous section, we turn our attention now to identifying the best-fitting survival sub-model (i.e. ignoring the longitudinal nature of the data), which will be employed in *all* subsequent joint models, as well as the best-fitting GLMM for each of the longitudinal responses we consider.

7.3.1 The survival sub-model

In Table 7.1 we tabulated the covariates we consider to be baseline for the purposes of analyses presented in this chapter, naming them (receipt of) `drug`, (standardised) `age`, `sex` (female) and `histologic` status. Since those with cirrhosis drastically inflate the hazard of mortality, which will likely be problematic, we combine the disease stages 0–1 and 3–4 together in a re-definition of `histologic` for analysis purposes. We note at the outset that we do *not* consider interactions between baseline covariates due to limitations with `gmvjoint`, which is used to fit all joint models in future sections. Further discussion of the limitations of `gmvjoint` is provided in Appendix D.4.

We begin with the (saturated) four-variate model which we present in Appendix B.3.1. The covariates `age` and `histologic` appear to be very significantly associated with mortality, with being female also significantly protective at the 5% level. Recipiency of the study drug (as we may expect from Table 7.1) does not appear to hold association here. There

is evidence to remove the study drug and re-fit the PH model, then. Doing so, we fit the trivariate Cox PH model

$$\lambda_i(t) = \lambda_0(t) \exp\{\texttt{age} \times \zeta_1 + \texttt{sex} \times \zeta_2 + \texttt{histologic} \times \zeta_3\}, \tag{7.1}$$

with resultant parameter estimates presented in Table 7.2

Parameter	Estimate	exp{Estimate}	Standard Error	Z	p-value
ζ_1	0.41	1.50	0.09	4.55	< 0.001
ζ_2	-0.45	0.64	0.22	-2.00	0.049
ζ_3	0.99	2.70	0.23	4.24	< 0.001

Table 7.2: Parameter estimates, presented with their standard errors and exponentiated value for the Cox PH model fit to PBC data using all considered baseline covariates bar `drug`.

As a final consideration, we fit every possible trivariate, bivariate, and univariate Cox PH model and collate the resultant values for AIC (defined in the same way as (6.5) in Section 6.2) and Harrell's C-index, which is a goodness-of-fit measure commonly used in survival analysis (higher values better). These two criteria are presented in Figure 7.3 for each model; we note the model we considered in (7.1) attains the lowest AIC and amongst the highest C-index. The (further) reduced model containing only `age` and `histologic` we note comes close in both measures, but the inclusion of `sex` is borderline significant (test statistic $\chi^2 = 3.678$, p-value=0.055) and so is included going forward.

We can therefore be confident that the model we arrive to here is the 'best available' given the available covariates. Since `histologic` and `age` are 'most associated' in Table 7.2, we are compelled to monitor attenuation in estimates for $\boldsymbol{\zeta}$ in presence of longitudinal biomarkers in subsequent joint models. Hereafter, every survival sub-model in Section 7.4 takes the form (7.1).

7.3.2 Longitudinal sub-models

We now turn attention to the more complex issue of identifying the best-fitting model for *each* of the longitudinal biomarkers we identified in Section 7.2. Since these are temporal in nature, we also need to consider the time specification (e.g. linear, quadratic, and so on) in addition to the baseline covariates which we considered in the survival model selection in the previous section.

In an effort to streamline the process somewhat, we model (log) serum bilirubin, (log) AST and serum albumin to be Gaussian in keeping with existing literature (to name but two, Hickey et al. (2018a) and Rustand et al. (2023)). However, in a departure from said existing literature, we use the Gamma distribution for prothrombin time since in

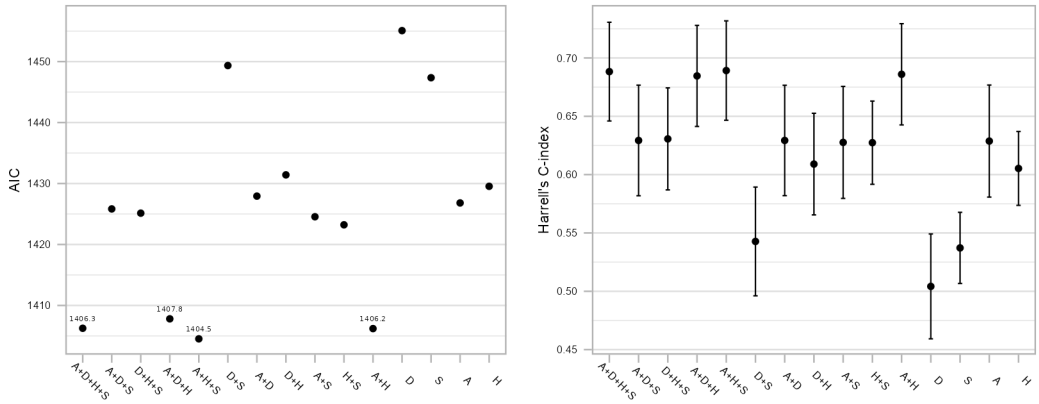


Figure 7.3: AIC and Harrell's C-index for each possible trivariate, bivariate, and univariate Cox model fit to available covariates. `A`: `age`; `D`: `drug`; `H`: `histologic`; `S`: `sex`. The four lowest AIC values are superimposed.

Section 4.3.3 we explicitly noted this family's use for modelling times. We utilise the negative binomial and generalised Poisson for alkaline and platelet count since we note these produced much better model fits than the regular Poisson family (results not shown).

For each of the longitudinal biomarkers, we (attempt to) fit every possible GLMM constructed by a possibly linear, quadratic, or natural cubic splines (with knots at tertiles of follow-up) `time` specification, with some combination of `drug`, `age`, `sex` and `histologic` as defined previously. For each `time` specification we elect a random effects structure which mimics the chosen temporal structure i.e. if `time` is to be modelled as quadratic in the fixed effects then the random effects are defined as $b_{i0} + b_{i1}\texttt{time} + b_{i2}\texttt{time}^2$. For the linear `time` model we separately consider a random intercept only specification. In their application to PBC data Rustand et al. (2023) elect a `drug` interaction with `time` which we additionally consider in each possible fit, too.

Fitting each model by `glmmTMB` (Brooks et al., 2017), we collect the summary measures of AIC, BIC, and the log-likelihood of each successfully fitted model. If a singular fit is returned by `glmmTMB` – indicating that the model is overfit – or else the model fit is unsuccessful then the corresponding model formula is discarded and not considered further. As we mentioned in Section 6.2 these information criteria often return different models. Since BIC penalises model complexity much more harshly, we identify the best fitting by this criterion in an effort to favour the more parsimonious model in each case.

The resultant BICs are presented in Figure 7.4. Here we observe that a mixture of linear, quadratic, and natural cubic spline time specifications are chosen. No drug interaction is present in any model identified best fitting for any response. The `histologic` status is present in each model, in combination with a mixture of either `age` or `sex`; interestingly

no model is chosen with more than two baseline covariates included.

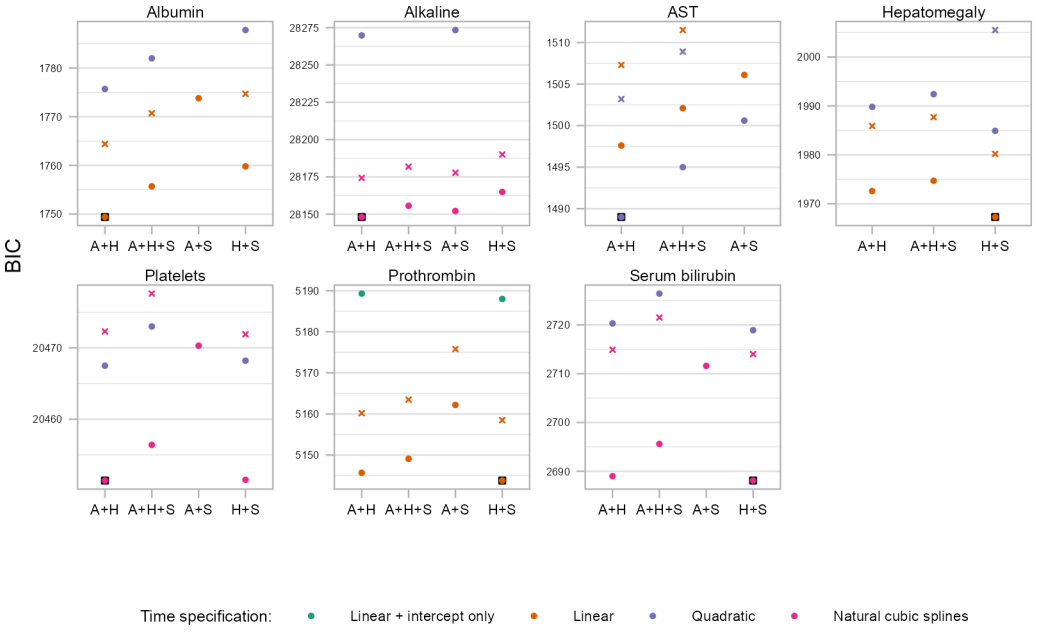


Figure 7.4: BIC values for each longitudinal response. A cross ('x') indicates `drug-time` interaction and a closed circle no interaction. The x-axis presents the combination of baseline covariates `A`: `age`; `H`: `histologic`; `S`: `sex`. A black square is drawn behind the model with the lowest BIC for each response for clarity's sake.

With the best longitudinal model identified by this long-winded supervised process, we have the additional task of electing dispersion models for those applicable models: Alkaline (negative binomial); prothrombin time (Gamma) and platelet count (generalised Poisson). To continue along the more parsimonious line of thinking – reflected in our prior choice to employ BIC – we consider *only five* alternative (to the global intercept already in place thus far) univariate dispersion models: Each of the four baseline covariates and `time`. The results from this exercise are presented in Appendix C.3.2. We take forward a dispersion model of `time` for both platelets and prothrombin time, and `histologic` for alkaline.

To bring this exercise to a close, we present all chosen models for clarity's sake in Table 7.3

7.4 Joint modelling

In the previous section, we identified what we believe to be the 'best performing' survival sub-model (7.1) – which we don't change in any subsequent fit unless stated – as well as the most parsimonious GLMM for each of the seven biomarkers given in Table 7.3. We

	Longitudinal model				
Response	Time specification	Other covariates	Dispersion model	P	q
Albumin	Linear	`age`, `histologic`	N/A	4	2
Alkaline	Natural cubic splines	`age`, `histologic`	`histologic`	8	4
AST	Quadratic	`age`, `histologic`	N/A	5	3
Hepatomegaly	Linear	`histologic`, `sex`	N/A	4	2
Platelet count	Natural cubic splines	`age`, `histologic`	`time`	8	4
Prothrombin time	Linear	`histologic`, `sex`	`time`	6	2
Serum bilirubin	Natural cubic splines	`histologic`, `sex`	N/A	6	4

Table 7.3: Chosen models for each longitudinal response we consider in the PBC application. The number of parameters (determined by the longitudinal *and* dispersion models) is given by P, notably the number of covariance parameters vech(D) is *excluded*, and the dimension of random effects by q.

now move on to the joint modelling of the longitudinal and time-to-event processes. We begin by establishing association with mortality on a biomarker-by-biomarker basis before considering multivariate fits.

7.4.1 Univariate joint models

We are predominantly interested in the estimated survival parameters $\hat{\boldsymbol{\Phi}} = \left(\hat{\gamma}, \hat{\boldsymbol{\zeta}}^\top\right)^\top$ obtained from each of seven univariate joint models (one for each biomarker). Before proceeding we recall the parameter estimates for $\hat{\boldsymbol{\zeta}} = \left(\hat{\zeta}_1, \hat{\zeta}_2, \hat{\zeta}_3\right)^\top$ we obtained when considering only the baseline covariates `age`, `sex` and `histologic` in Table 7.2, noting attenuation (if any) that occur when each biomarker is jointly modelled.

The survival parameter estimates in Figure 7.5 reveal univariate association between the levels of each biomarker, given by association parameter estimate $\hat{\gamma}$, and mortality. The estimates for the association parameter indicate that increased levels (in the linear predictor) of alkaline phosphotase; AST; hepatomegaly; prothrombin time and serum bilirubin all *increase* the hazard, the same being true for lower levels of platelet count and serum albumin. Focussing specifically on serum bilirubin, the point estimate [95% CI] 1.36 [1.20, 1.52] allows us to infer that individuals whose levels of serum bilirubin at some time t are one unit higher than the population average (i.e. due to $\boldsymbol{b}_i$) tend to experience 36% higher risk of mortality.

A very large value for $\hat{\gamma}$ is observed in the prothrombin joint model (point estimate [95% CI] 12.70 [10.10, 15.31]), and is scaled proportionally by the average value of the estimated random effects in Figure 7.5. We note the unscaled estimate is not worlds away from the absolute value observed in an analogous Gaussian fit conducted by Hickey et al. (2018a) on a Box-Cox transformed prothrombin, the large value here stemming from relatively small random effects.

In each case, taking account of the longitudinal biomarker does not appear to greatly

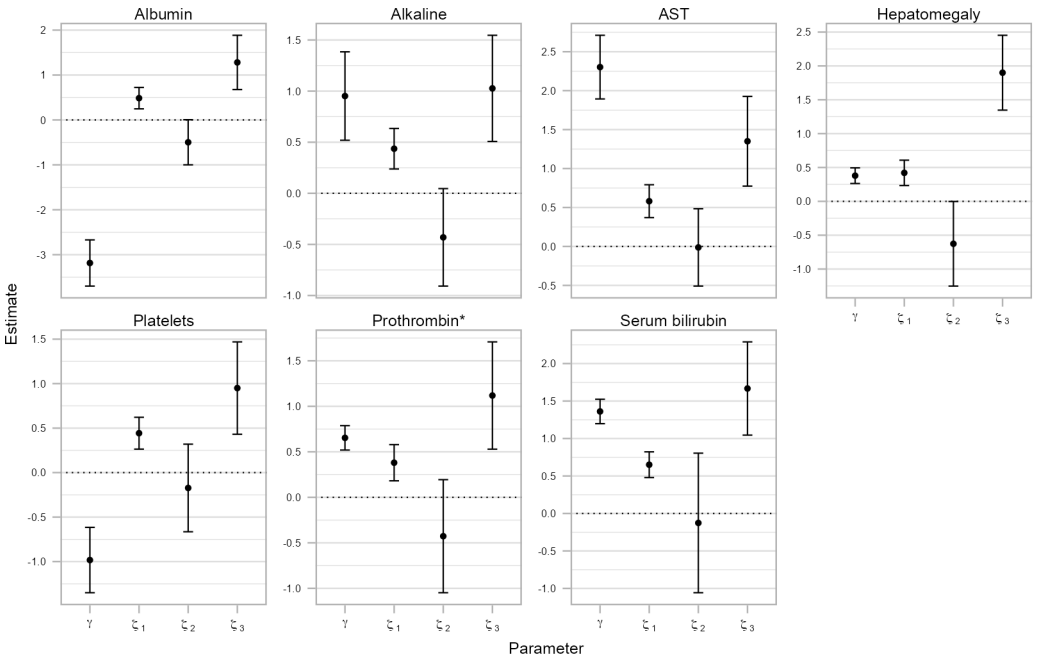


Figure 7.5: Survival parameter estimates $\hat{\boldsymbol{\Phi}}$ (with 95% confidence interval shown) from univariate joint models fit to each biomarker (denoted by panel title). The longitudinal and dispersion sub-models are given in Table 7.3 and the survival sub-model by (7.1). Prothrombin is asterisked as its $\hat{\gamma}$ term is scaled by 0.05 for presentation, which is approximately the 80^{th} percentile of estimated random effects at $t = 1$.

attenuate the estimate for $\hat{\zeta}_1$ (`age`), with the same largely being true for $\hat{\zeta}_3$ (`histologic`), besides for serum bilirubin and hepatomegaly, where the point estimate is increased. Interestingly, the biomarker's presence in the survival sub-model leads to $\hat{\zeta}_2$ (`sex`) – which only held borderline significance in the standalone model in Table 7.2 – largely holding either no, or still only borderline, significance for the most part. Note however that we do not undertake model reduction and re-fitting on each univariate joint model fit here, opting instead to perform this when considering the multivariate case in the next sections.

The Pearson residuals (derived in the same manner as presented in Section 6.1.1) are presented in Appendix C.3.3. We note little awry amongst these results; save for some larger residuals for, say, alkaline. Two exemplar Cox-Snell residual plots are shown in Appendix C.3.4, indicating broadly good agreement with observed and theoretical curves.

7.4.2 Multivariate joint models

When considering how best to build a multivariate joint model, we first revisit prior discussion and consternation from Appendix A.3 and Sections 3.4.3 and 3.5 that a large number of longitudinal responses being jointly modelled likely requires a larger sample

size or incidence of failure, which would perhaps admonish against simply fitting one seven-variate model.

With this in mind then, we consider a joint model which attempts to group the biomarkers into broad categories of liver function, such that those we expect to be *most correlated* with one another (since ostensibly they measure the same, or similar, pathological states) are first modelled together, in an effort to establish the 'more dominant' biomarker(s) (determined by the observed association(s) with mortality), which we then carry forward.

Specifically, we consider three non-overlapping multivariate joint models, each conveniently consisting of biomarkers of differing types. We group together prothrombin time and platelet count into '*Blood clotting and flow*'; alkaline, AST, and serum bilirubin into '*Liver enzymes*'; and finally hepatomegaly and serum albumin into '*Liver health and function*'.

The parameter estimates for $\hat{\mathbf{\Phi}}$ obtained from these intermediary multivariate joint models are presented in Figure 7.6. Here we note that for the '*Blood clotting and flow*' model, prothrombin retains its large association parameter $\hat{\gamma}$, and diminishes the standalone association we observed for platelet count in Figure 7.5. Similar phenomena occurs in '*Liver enzymes*', where serum bilirubin dominates alkaline and AST, and in '*Liver health and function*' in the presence of serum albumin, hepatomegaly is no longer considered significant, albumin explaining the purported association we observed at the univariate stage here; this interpretation extends to the other observed attenuations.

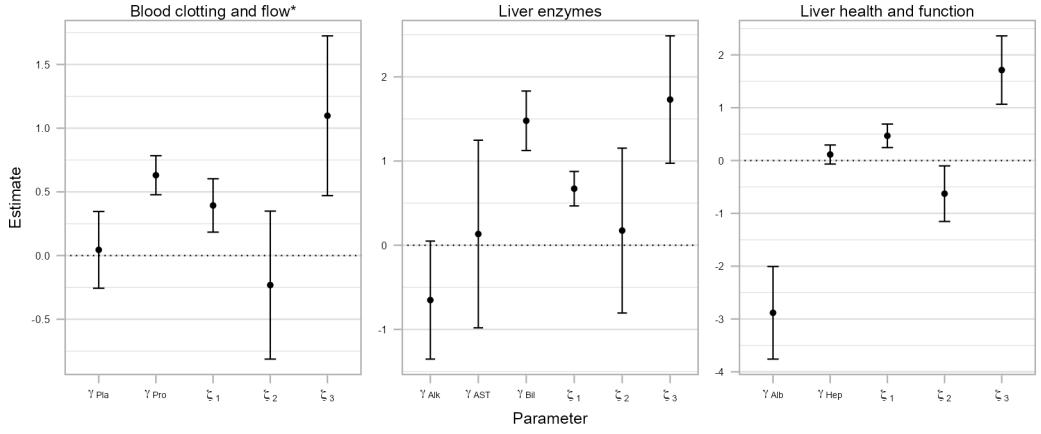


Figure 7.6: Survival parameter estimates $\hat{\mathbf{\Phi}}$ (with 95% confidence interval shown) from the intermediate multivariate joint models with grouped biomarkers (denoted by panel title). The subscripts for $\hat{\gamma}$ contain the first three letters of the corresponding biomarker to facilitate identification here. The longitudinal and dispersion sub-models are given in Table 7.3 and the survival sub-model by (7.1). Blood clotting and flow is asterisked as the presented $\hat{\boldsymbol{\gamma}}$ are scaled by $(0.50, 0.05)$ for presentation, which is approximately the 80^{th} percentile of estimated random effects at $t = 1$ for each biomarker.

The full parameter estimates alongside the estimated covariance matrices (with correla-

tions) for each model are presented in Appendix B.3.2. The correlations here elucidate the relationships between e.g. having higher values, or say a steeper trajectory, in one biomarker and another. Perhaps being of particular note is the relationship (i.e. high correlation) between trajectories of AST and serum bilirubin in the trivariate '*Liver enzymes*' fit. The available information on the hazard here is correlated, and we don't 'need' both biomarkers; the signal from serum bilirubin is stronger than that from AST in this instance.

These three joint models point us toward the next stage of model building: A trivariate joint model consisting of (log) serum bilirubin, albumin and prothrombin time.

7.5 Arriving at the final model

At the end of the multivariate joint model building exercise, we whittled down three joint models which broadly pertained to certain categories of liver function each to a sole longitudinal response which held the greatest association with mortality. In this section we bring the process to a close.

We begin with the trivariate model we previously identified. Specifically, we fit the model

$$
\begin{cases}
\begin{cases}
\log(\mathbb{E}[\text{Prothrombin time}|\boldsymbol{b}_{i1}]) & = (\beta_{10} + b_{i10}) + (\beta_{11} + b_{i11})\, t \\
& \quad + \texttt{H}_i \times \beta_{12} + \texttt{S}_i \times \beta_{13} \\
\log(\boldsymbol{\varphi}_{i1}) & = \sigma_{10} + \sigma_{11} t
\end{cases} \\
\begin{cases}
\log(\text{Serum bilirubin}) & = (\beta_{20} + b_{i20}) + (\beta_{21} + b_{i21})\, N_1(t) \\
& \quad + (\beta_{22} + b_{i22})\, N_2(t) + (\beta_{23} + b_{i23})\, N_3(t) \\
& \quad + \texttt{H}_i \times \beta_{24} + \texttt{S}_i \times \beta_{25} + \varepsilon_{i2}(t)
\end{cases} \\
\left\{ \text{Albumin} \quad = (\beta_{30} + b_{i30}) + (\beta_{31} + b_{i31})\, t + \ \texttt{A}_i \times \beta_{32} + \texttt{H}_i \times \beta_{33} + \varepsilon_{i3}(t) \right. \\
\lambda_i(t) = \lambda_0(t) \exp\left\{ \texttt{A}_i \times \zeta_1 + \texttt{S}_i \times \zeta_2 + \texttt{H}_i \times \zeta_3 + \sum_{k=1}^{3} \gamma_k \boldsymbol{W}_k(t)^\top \boldsymbol{b}_{ik} \right\},
\end{cases}
\tag{7.2}
$$

where subject i's `age`, `sex` and `histologic` state is given by $\texttt{A}_i$, $\texttt{S}_i$ and $\texttt{H}_i$ respectively and $N_1(t), \ldots, N_3(t)$ denotes the set of natural cubic splines with knots at tertiles of follow-up. Left braces are used to separate each sub-model (and its dispersion model) visually, with the left-most brace denoting that these are to be jointly modelled.

The results from this trivariate model are presented in Table 7.4. Alongside the results obtained using the methodology outlined in Chapters 3 and 4 we additionally present results from an analogous model fit using conventional software `JMbayes2` (Rizopoulos et al., 2021). We note at the outset that `JMbayes2` fits the joint model under a different

parameterisation of association; the *current value* of the linear predictor placed in (2.2) instead of the shared random effects. Additionally, a smoothed baseline hazard is estimated by `JMbayes2`, whereas it is unspecified in (2.2). Therefore, the results for $\hat{\boldsymbol{\gamma}}$ (and by extension $\hat{\boldsymbol{\zeta}}$) are not *directly* comparable. The MCMC fits are conducted with 2000 iterations of burnin and 10,000 iterations afterwards across three chains.

With this in mind, we note the results show broadly good agreement between the approximate method and the 'gold standard' of MCMC. The fixed effects $\hat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\beta}}_1^\top, \hat{\boldsymbol{\beta}}_2^\top, \hat{\boldsymbol{\beta}}_3^\top\right)^\top$ are all of similar magnitude and direction, with the same being true for $\text{vech}\left(\hat{\mathrm{D}}\right)$. Across all parameters we note that the MCMC approach yields parameter estimates with far less uncertainty around them; we discuss inherent differences between these methodological approaches later in Section 7.7.

Both approaches agree that prothrombin ($\hat{\gamma}_1$) is *not* significantly associated with mortality: The 95% confidence (and credible) intervals straddling zero. The signals provided by the other two biomarkers explain the purported association we previously noted. We note agreement in magnitude and direction for $\hat{\gamma}_2$ and $\hat{\gamma}_3$.

Focusing lastly on computation time, the time taken for the approximate EM algorithm to converge and standard errors be calculated was 55.104 seconds, which (unsurprisingly) is approximately a 4.5-fold decrease in the computation time for the `JMbayes2` fit. The total computation time (i.e. including obtention of initial conditions) for approximate EM was 59.993 seconds. We note the comparison in elapsed time is largely unfair; simply emphasising the ability to obtain a comparable set of parameter estimates in a relatively fast manner.

There is strong evidence in Table 7.4 – given the discussion above – that prothrombin should be removed and the model reduced to a bivariate joint model containing only (log) serum bilirubin and albumin. To be explicit, we now fit

$$\begin{cases} \text{log(Serum bilirubin)} & = (\beta_{10} + b_{i10}) + (\beta_{11} + b_{i11})\, N_1(t) \\ & \quad + (\beta_{12} + b_{i12})\, N_2(t) + (\beta_{13} + b_{i13})\, N_3(t) \\ & \quad + \texttt{H}_i \times \beta_{14} + \texttt{S}_i \times \beta_{15} + \varepsilon_{i1}(t) \\ \text{Albumin} & = (\beta_{20} + b_{i20}) + (\beta_{21} + b_{i21})\, t + \texttt{A}_i \times \beta_{22} \\ & \quad + \texttt{H}_i \times \beta_{23} + \varepsilon_{i2}(t) \\ \lambda_i(t) & = \lambda_0(t) \exp\left\{ \texttt{A}_i \times \zeta_1 + \texttt{S}_i \times \zeta_2 + \texttt{H}_i \times \zeta_3 \right. \\ & \qquad \left. + \sum_{k=1}^{2} \gamma_k \boldsymbol{W}_k(t)^\top \boldsymbol{b}_{ik} \right\}, \end{cases} \tag{7.3}$$

The resultant parameter estimates are presented in Table 7.5. Once more we present

	Approximate EM		`JMbayes2`	
Parameter	Estimate (SE)	95% CI	Mean (SD)	95% CrI
$D_{1,00}$	0.005 (0.001)	[0.004, 0.007]	0.006 (0.001)	[0.004, 0.008]
$D_{1,10}$	0.000 (0.000)	[0.000, 0.000]	0.000 (0.000)	[0.000, 0.000]
$D_{1,11}$	0.000 (0.000)	[0.000, 0.000]	0.000 (0.000)	[0.000, 0.000]
β_{10}	**2.344** (0.016)	[2.312, 2.376]	**2.356** (0.014)	[2.329, 2.383]
β_{11}	**0.017** (0.001)	[0.015, 0.020]	**0.017** (0.001)	[0.014, 0.020]
β_{12}	**0.045** (0.013)	[0.020, 0.070]	**0.044** (0.010)	[0.025, 0.063]
β_{13}	-0.024 (0.014)	[-0.051, 0.003]	**-0.027** (0.012)	[-0.050, -0.003]
σ_{10}	**4.893** (0.039)	[4.817, 4.969]	**5.249** (0.048)	[5.172, 5.324]
σ_{11}	**0.124** (0.012)	[0.101, 0.147]		
γ_1	1.754 (3.659)	[-5.417, 8.925]	2.389 (2.012)	[-1.700, 6.338]
$D_{2,00}$	0.905 (0.100)	[0.709, 1.101]	0.912 (0.076)	[0.775, 1.073]
$D_{2,10}$	0.171 (0.173)	[-0.167, 0.510]	0.148 (0.125)	[-0.101, 0.399]
$D_{2,20}$	0.477 (0.280)	[-0.071, 1.025]	0.285 (0.142)	[0.009, 0.599]
$D_{2,30}$	0.459 (0.407)	[-0.339, 1.257]	0.195 (0.216)	[-0.233, 0.692]
$D_{2,11}$	1.789 (0.348)	[1.106, 2.471]	1.765 (0.287)	[1.260, 2.407]
$D_{2,21}$	1.775 (0.429)	[0.935, 2.616]	1.425 (0.215)	[1.027, 1.876]
$D_{2,31}$	0.820 (0.472)	[-0.105, 1.745]	0.398 (0.255)	[-0.121, 0.857]
$D_{2,22}$	2.638 (0.768)	[1.132, 4.144]	2.110 (0.356)	[1.467, 2.842]
$D_{2,32}$	1.556 (0.837)	[-0.084, 3.197]	1.017 (0.342)	[0.430, 1.744]
$D_{2,33}$	1.906 (0.807)	[0.325, 3.487]	1.604 (0.379)	[0.987, 2.444]
β_{20}	0.218 (0.220)	[-0.213, 0.649]	**0.342** (0.171)	[0.008, 0.681]
β_{21}	**1.070** (0.139)	[0.797, 1.342]	**1.168** (0.125)	[0.923, 1.417]
β_{22}	**1.485** (0.210)	[1.075, 1.896]	**1.484** (0.156)	[1.196, 1.816]
β_{23}	**1.452** (0.327)	[0.812, 2.093]	**1.302** (0.242)	[0.865, 1.829]
β_{24}	**0.608** (0.157)	[0.301, 0.915]	**0.595** (0.124)	[0.347, 0.838]
β_{25}	-0.212 (0.190)	[-0.585, 0.160]	-0.268 (0.149)	[-0.558, 0.025]
σ_2^2	0.076 (0.002)	[0.072, 0.081]	0.076 (0.003)	[0.070, 0.083]
γ_2	**1.067** (0.163)	[0.747, 1.387]	**1.124** (0.140)	[0.848, 1.398]
$D_{3,00}$	0.110 (0.015)	[0.081, 0.139]	0.109 (0.012)	[0.087, 0.134]
$D_{3,10}$	0.002 (0.003)	[-0.003, 0.007]	0.002 (0.002)	[-0.002, 0.005]
$D_{3,11}$	0.003 (0.001)	[0.001, 0.005]	0.002 (0.001)	[0.002, 0.004]
β_{30}	**3.722** (0.056)	[3.612, 3.832]	**3.697** (0.039)	[3.619, 3.773]
β_{31}	**-0.093** (0.006)	[-0.105, -0.081]	**-0.095** (0.006)	[-0.107, -0.084]
β_{32}	**-0.058** (0.025)	[-0.108, -0.009]	**-0.063** (0.018)	[-0.097, -0.029]
β_{33}	**-0.212** (0.064)	[-0.337, -0.086]	**-0.209** (0.044)	[-0.296, -0.123]
σ_3^2	**0.097** (0.002)	[0.094, 0.101]	**0.098** (0.004)	[0.091, 0.106]
γ_3	-1.912 (0.711)	[-3.306, -0.519]	-2.158 (0.471)	[-3.169, -1.306]
ζ_1	**0.559** (0.115)	[0.335, 0.784]	**0.459** (0.134)	[0.196, 0.722]
ζ_2	-0.474 (0.433)	[-1.323, 0.376]	-0.252 (0.296)	[-0.828, 0.329]
ζ_3	**1.631** (0.447)	[0.756, 2.506]	0.528 (0.272)	[-0.004, 1.079]

Table 7.4: Parameter estimates (SE) for the trivariate model (7.2) applied to the PBC data. Total computation time for the approximate EM algorithm was 59.993 seconds. Horizontal dashed lines are used to visually separate results for each response, and the time-invariant survival parameters $\hat{\boldsymbol{\zeta}}$ which aren't associated with a specific response and reported separately. Parameter estimates (SD) from `JMbayes2` are additionally reported ('CrI': Credible interval). Computation time for the MCMC scheme in the `JMbayes2` fit was 283.329 seconds, not including time spent in obtention of its initial conditions. `JMbayes2` does not allow for dispersion sub-models, and so σ_{11} is left blank here. The variance-covariance matrix $\hat{\mathrm{D}}_k$ is reported for each of the responses in the form $\hat{\mathrm{D}}_{k,ef}$ where k denotes the longitudinal response, and e, f the random effect indices. **Bold** parameter estimates indicate statistical significance at the 5% level for ease of visual comparison across approaches.

parameter estimates from an analogous fit by `JMbayes2` to act as a 'barometer'. Since now $\boldsymbol{Y}_{ik}|\boldsymbol{b}_{ik} \sim \mathrm{MVN}(\cdot)\,,\ k = 1, 2$ we additionally present parameter estimates from a multivariate joint model fit by `joineRML` (Hickey et al., 2018a), utilising the same model control arguments here as in Section 3.5.

Comparing the elapsed time for convergence to be achieved by the two ML approaches (i.e. discounting initial conditions etc.), the approximate EM algorithm converged in 33.5 seconds, and `joineRML` in 53.6, indicating that the 'flattening' of multidimensional integrals

($q = 6$) leads to approximately a 40% decrease in time spent in the algorithm. Indeed, in Section 3.5 we observed that this gain in performance only increases with q.

As we mentioned in Section 3.5 we expect very good agreement in parameter estimates across the approximate EM algorithm and `joineRML` since the approaches to the modelling process and model parameterisation is the same, in comparison with the Bayesian `JMbayes2`'s slightly different parameterisation of the joint model already discussed.

Inspecting the parameter estimates across approaches, we observe excellent agreement between the approximate EM and the established ML approach `joineRML`. When incorporating `JMbayes2` into the inspection we note that some inferential differences do exist, for instance the MCMC approach finds being female to significantly reduce the (log) value of serum bilirubin, whereas the two maximum likelihood approaches do not. The association parameters indicate higher bilirubin values and lower albumin values increase the log hazard.

Overall, we observe good agreement in both sign and magnitude of $\hat{\boldsymbol{\Omega}}$ across approaches. The largest discrepancy appears to be for $\hat{\zeta}_3$ (the coefficient attached to `histologic`), where `JMbayes2` declares it to hold only borderline significance. This may be due to the alternative parameterisation of the survival sub-model, or the prior imposed on the survival parameters perhaps 'anchoring' the estimates towards the null. All approaches declare sex to not be associated with mortality.

7.6 A final model and post-hoc analyses

Across Sections 7.4 and 7.5 we have built-up, and whittled down, to a bivariate joint model of (log) serum bilirubin and serum albumin. In the last model we presented in the previous section (7.3), we noted there is evidence (across *all* modelling approaches) to remove `sex` from the survival sub-model as seen in Table 7.5. Indeed, in the first instance in Section 7.3.1 we noted only borderline significance (see Table 7.2), and upon inclusion of the longitudinal responses its coefficient quickly attenuated toward the null (e.g. in Figures 7.5 and 7.6). Removing this covariate from the survival sub-model, we arrive at

	Approximate EM		`JMbayes2`		`joineRML`	
Parameter	Estimate (SE)	95% CI	Mean (SD)	95% CrI	Estimate (SE)	95% CI
$D_{1,00}$	0.905 (0.098)	[0.714, 1.097]	0.921 (0.109)	[0.732, 1.154]	0.904 (0.097)	[0.714, 1.095]
$D_{1,10}$	0.127 (0.161)	[-0.188, 0.443]	0.098 (0.133)	[-0.162, 0.369]	0.106 (0.154)	[-0.196, 0.408]
$D_{1,20}$	0.403 (0.230)	[-0.049, 0.855]	0.302 (0.156)	[0.012, 0.648]	0.400 (0.224)	[-0.039, 0.839]
$D_{1,30}$	0.404 (0.340)	[-0.263, 1.071]	0.294 (0.256)	[-0.203, 0.870]	0.432 (0.327)	[-0.208, 1.073]
$D_{1,11}$	1.703 (0.312)	[1.091, 2.314]	1.727 (0.277)	[1.265, 2.357]	1.597 (0.292)	[1.025, 2.169]
$D_{1,21}$	1.646 (0.366)	[0.928, 2.364]	1.377 (0.226)	[0.968, 1.853]	1.662 (0.352)	[0.972, 2.351]
$D_{1,31}$	0.763 (0.375)	[0.029, 1.497]	0.419 (0.264)	[-0.144, 0.940]	0.834 (0.350)	[0.149, 1.519]
$D_{1,22}$	2.440 (0.600)	[1.264, 3.616]	2.080 (0.342)	[1.474, 2.839]	2.420 (0.579)	[1.285, 3.556]
$D_{1,32}$	1.464 (0.651)	[0.187, 2.740]	1.108 (0.349)	[0.506, 1.950]	1.469 (0.619)	[0.255, 2.683]
$D_{1,33}$	1.882 (0.665)	[0.578, 3.186]	1.735 (0.425)	[1.024, 2.749]	1.882 (0.649)	[0.610, 3.154]
β_{10}	0.246 (0.209)	[-0.163, 0.655]	**0.368** (0.170)	[0.033, 0.703]	0.299 (0.209)	[-0.111, 0.709]
β_{11}	**1.052** (0.128)	[0.801, 1.303]	**1.121** (0.124)	[0.881, 1.369]	**1.064** (0.120)	[0.829, 1.300]
β_{12}	**1.450** (0.159)	[1.138, 1.762]	**1.502** (0.148)	[1.231, 1.816]	**1.517** (0.153)	[1.218, 1.816]
β_{13}	**1.420** (0.250)	[0.929, 1.911]	**1.399** (0.223)	[1.004, 1.877]	**1.509** (0.242)	[1.035, 1.983]
β_{14}	**0.602** (0.150)	[0.309, 0.896]	**0.601** (0.124)	[0.359, 0.844]	**0.625** (0.147)	[0.336, 0.914]
β_{15}	-0.228 (0.178)	[-0.577, 0.122]	**-0.302** (0.150)	[-0.596, -0.006]	-0.285 (0.184)	[-0.644, 0.075]
σ_1^2	0.076 (0.002)	[0.072, 0.080]	0.076 (0.003)	[0.070, 0.083]	0.077 (0.002)	[0.073, 0.081]
γ_1	**1.082** (0.145)	[0.798, 1.367]	**1.148** (0.133)	[0.893, 1.411]	**1.117** (0.148)	[0.828, 1.407]
$D_{2,00}$	0.111 (0.014)	[0.083, 0.138]	0.109 (0.012)	[0.087, 0.135]	0.108 (0.014)	[0.081, 0.135]
$D_{2,10}$	0.001 (0.002)	[-0.004, 0.005]	0.001 (0.002)	[-0.003, 0.005]	0.001 (0.002)	[-0.003, 0.006]
$D_{2,11}$	0.003 (0.001)	[0.001, 0.004]	0.003 (0.001)	[0.002, 0.004]	0.003 (0.001)	[0.001, 0.004]
β_{20}	**3.723** (0.053)	[3.620, 3.826]	**3.707** (0.040)	[3.628, 3.785]	**3.719** (0.052)	[3.617, 3.822]
β_{21}	**-0.091** (0.006)	[-0.102, -0.080]	**-0.094** (0.006)	[-0.106, -0.084]	**-0.093** (0.005)	[-0.104, -0.083]
β_{22}	**-0.078** (0.022)	[-0.122, -0.034]	**-0.082** (0.018)	[-0.117, -0.046]	**-0.082** (0.022)	[-0.126, -0.038]
β_{23}	**-0.219** (0.061)	[-0.339, -0.100]	**-0.225** (0.046)	[-0.315, -0.134]	**-0.229** (0.061)	[-0.348, -0.110]
σ_2^2	0.097 (0.002)	[0.094, 0.100]	0.098 (0.004)	[0.091, 0.106]	0.097 (0.002)	[0.094, 0.101]
γ_2	**-2.178** (0.463)	[-3.085, -1.271]	**-2.505** (0.428)	[-3.315, -1.690]	**-2.360** (0.459)	[-3.260, -1.459]
ζ_1	**0.597** (0.108)	[0.386, 0.808]	**0.432** (0.133)	[0.174, 0.695]	**0.627** (0.109)	[0.412, 0.841]
ζ_2	-0.496 (0.390)	[-1.261, 0.268]	-0.314 (0.292)	[-0.871, 0.257]	-0.586 (0.409)	[-1.386, 0.215]
ζ_3	**1.654** (0.432)	[0.807, 2.502]	**0.577** (0.271)	[0.066, 1.131]	**1.800** (0.445)	[0.927, 2.672]

Table 7.5: Parameter estimates (SE) for the bivariate model (7.3) applied to the PBC data. Total computation time for the approximate EM algorithm was 36.618 seconds. Horizontal dashed lines are used to visually separate results for each response, and the time-invariant survival parameters $\hat{\boldsymbol{\zeta}}$ which aren't associated with a specific response and reported separately. Parameter estimates (SD) from `JMbayes2` are additionally reported ('CrI': Credible interval) along with `joineRML`. Computation time for the MCMC scheme in the `JMbayes2` fit was 148.458 seconds, not including time spent in obtention of its initial conditions; total computation time for `joineRML` was 80.090 seconds. The variance-covariance matrix $\hat{\mathrm{D}}_k$ is reported for each of the responses in the form $\hat{\mathrm{D}}_{k,ef}$ where k denotes the longitudinal response, and e, f the random effect indices. **Bold** parameter estimates indicate statistical significance at the 5% level for ease of visual comparison across approaches.

our 'final' model

$$
\begin{cases}
\log(\text{Serum bilirubin}) & = (\beta_{10} + b_{i10}) + (\beta_{11} + b_{i11})\, N_1(t) \\
& \quad + (\beta_{12} + b_{i12})\, N_2(t) + (\beta_{13} + b_{i13})\, N_3(t) \\
& \quad + \mathtt{H}_i \times \beta_{14} + \mathtt{S}_i \times \beta_{15} + \varepsilon_{i1}(t) \\
\text{Albumin} & = (\beta_{20} + b_{i20}) + (\beta_{21} + b_{i21})\, t + \mathtt{A}_i \times \beta_{22} \\
& \quad + \mathtt{H}_i \times \beta_{23} + \varepsilon_{i2}(t) \\
\lambda_i(t) & = \lambda_0(t) \exp\Big\{ \mathtt{A}_i \times \zeta_1 + \mathtt{H}_i \times \zeta_2 \\
& \qquad + \textstyle\sum_{k=1}^{2} \gamma_k \boldsymbol{W}_k(t)^\top \boldsymbol{b}_{ik} \Big\}.
\end{cases}
\tag{7.4}
$$

We once more provide tabulated results in Table 7.6, where we once more note good agreement across the available methods. The MCMC approach took approximately 160 seconds for its sampling to be completed, the approximate EM converged and calculated standard errors in 29.310 seconds, quite inexplicably this elapsed time for `joineRML` was 148.811 seconds[1].

	Approximate EM		`JMbayes2`		`joineRML`	
Parameter	Estimate (SE)	95% CI	Mean (SD)	95% CrI	Estimate (SE)	95% CI
$D_{1,00}$	0.905 (0.096)	[0.716, 1.094]	0.922 (0.111)	[0.735, 1.166]	0.903 (0.096)	[0.715, 1.090]
$D_{1,10}$	0.126 (0.161)	[-0.189, 0.441]	0.094 (0.129)	[-0.153, 0.351]	0.111 (0.154)	[-0.192, 0.413]
$D_{1,20}$	0.404 (0.230)	[-0.047, 0.854]	0.275 (0.140)	[0.017, 0.566]	0.414 (0.224)	[-0.025, 0.852]
$D_{1,30}$	0.402 (0.340)	[-0.264, 1.067]	0.254 (0.206)	[-0.095, 0.690]	0.446 (0.325)	[-0.192, 1.084]
$D_{1,11}$	1.702 (0.311)	[1.092, 2.311]	1.629 (0.278)	[1.128, 2.240]	1.603 (0.293)	[1.029, 2.178]
$D_{1,21}$	1.638 (0.363)	[0.926, 2.349]	1.364 (0.207)	[0.967, 1.777]	1.670 (0.353)	[0.979, 2.361]
$D_{1,31}$	0.753 (0.372)	[0.025, 1.481]	0.475 (0.237)	[0.022, 0.977]	0.853 (0.350)	[0.168, 1.538]
$D_{1,22}$	2.430 (0.598)	[1.258, 3.601]	2.077 (0.359)	[1.474, 2.882]	2.447 (0.583)	[1.303, 3.590]
$D_{1,32}$	1.452 (0.648)	[0.183, 2.722]	1.093 (0.357)	[0.553, 1.942]	1.499 (0.625)	[0.274, 2.723]
$D_{1,33}$	1.868 (0.661)	[0.572, 3.164]	1.650 (0.408)	[1.010, 2.664]	1.901 (0.658)	[0.611, 3.191]
β_{10}	0.184 (0.185)	[-0.179, 0.546]	**0.354** (0.172)	[0.016, 0.686]	0.218 (0.180)	[-0.134, 0.570]
β_{11}	**1.050** (0.128)	[0.799, 1.301]	**1.101** (0.122)	[0.867, 1.349]	**1.066** (0.120)	[0.831, 1.302]
β_{12}	**1.448** (0.158)	[1.138, 1.758]	**1.493** (0.157)	[1.200, 1.804]	**1.534** (0.152)	[1.235, 1.832]
β_{13}	**1.417** (0.249)	[0.929, 1.905]	**1.394** (0.245)	[0.940, 1.851]	**1.532** (0.240)	[1.061, 2.003]
β_{14}	**0.602** (0.150)	[0.309, 0.895]	**0.606** (0.125)	[0.360, 0.848]	**0.620** (0.147)	[0.332, 0.909]
β_{15}	-0.156 (0.135)	[-0.420, 0.108]	-0.292 (0.152)	[-0.588, 0.010]	-0.187 (0.131)	[-0.443, 0.070]
σ_1^2	0.076 (0.002)	[0.072, 0.080]	0.077 (0.003)	[0.070, 0.084]	0.077 (0.002)	[0.073, 0.081]
γ_1	**1.090** (0.141)	[0.813, 1.367]	**1.162** (0.134)	[0.907, 1.435]	**1.121** (0.144)	[0.838, 1.404]
$D_{2,00}$	0.110 (0.014)	[0.083, 0.137]	0.109 (0.012)	[0.088, 0.134]	0.108 (0.014)	[0.081, 0.135]
$D_{2,10}$	0.001 (0.002)	[-0.004, 0.005]	0.001 (0.002)	[-0.003, 0.004]	0.001 (0.002)	[-0.003, 0.006]
$D_{2,11}$	0.003 (0.001)	[0.001, 0.004]	0.002 (0.000)	[0.002, 0.004]	0.003 (0.001)	[0.001, 0.004]
β_{20}	**3.722** (0.053)	[3.619, 3.826]	**3.707** (0.040)	[3.629, 3.786]	**3.718** (0.052)	[3.615, 3.821]
β_{21}	**-0.091** (0.006)	[-0.102, -0.080]	**-0.093** (0.005)	[-0.104, -0.083]	**-0.093** (0.005)	[-0.104, -0.083]
β_{22}	**-0.077** (0.022)	[-0.120, -0.034]	**-0.081** (0.018)	[-0.117, -0.045]	**-0.081** (0.022)	[-0.124, -0.037]
β_{23}	**-0.220** (0.061)	[-0.339, -0.101]	**-0.226** (0.046)	[-0.317, -0.136]	**-0.228** (0.061)	[-0.348, -0.109]
σ_2^2	0.097 (0.002)	[0.094, 0.100]	0.098 (0.004)	[0.091, 0.106]	0.098 (0.002)	[0.094, 0.101]
γ_2	**-2.084** (0.445)	[-2.955, -1.212]	**-2.348** (0.409)	[-3.155, -1.556]	**-2.257** (0.441)	[-3.121, -1.392]
ζ_1	**0.637** (0.100)	[0.442, 0.833]	**0.476** (0.128)	[0.228, 0.726]	**0.669** (0.103)	[0.467, 0.871]
ζ_2	**1.637** (0.425)	[0.804, 2.470]	**0.570** (0.279)	[0.036, 1.131]	**1.759** (0.438)	[0.900, 2.619]

Table 7.6: Parameter estimates (SE) for the bivariate model (7.3) applied to the PBC data. Total computation time for the approximate EM algorithm was 31.656 seconds. Horizontal dashed lines are used to visually separate results for each response, and the time-invariant survival parameters $\hat{\boldsymbol{\zeta}}$ which aren't associated with a specific response and reported separately. Parameter estimates (SD) from `JMbayes2` are additionally reported ('CrI': Credible interval) along with `joineRML`. Computation time for the MCMC scheme in the `JMbayes2` fit was 159.739 seconds, not including time spent in obtention of its initial conditions; total computation time for `joineRML` was 175.013 seconds. The variance-covariance matrix $\hat{\mathrm{D}}_k$ is reported for each of the responses in the form $\hat{\mathrm{D}}_{k,ef}$ where k denotes the longitudinal response, and e, f the random effect indices. **Bold** parameter estimates indicate statistical significance at the 5% level for ease of visual comparison across approaches.

A graphical representation of the results in Table 7.6 are given in Figure 7.7. Several forms of model diagnostics (for the joint model fit by approximate EM only) are carried out. The plot of Pearson residuals, $\hat{\boldsymbol{r}}_1^{(P)}, \hat{\boldsymbol{r}}_2^{(P)}$, against fitted values (Appendix C.3.5), and QQ plots for the Pearson residuals and for the estimated random effects (Appendices C.3.7 and C.3.8, respectively), all allow us to infer that the usual modelling assumptions are met, but note some slight deviation away from the theoretical normal quantiles for Serum bilirubin. Additionally, posterior densities of these random effects (Appendix C.3.9)

[1]This strange result replicated across two machines, with a few attempts on each!

solidify this, though again some deviations may be occurring for the random intercept attached to serum bilirubin. Appraising the The Cox-Snell residuals (Appendix C.3.6 we note the survival function of the Cox-Snell residuals captures (roughly, at least at the 95% confidence level) the expected unit exponential. We can use the likelihood ratio test (6.4) to ensure that the removal of `sex` from the survival sub-model (7.3) to arrive at our 'final' model (7.4) is justified. We obtain LRT = 3.200 indicating only weak evidence (p-value=0.074) that we should include this extra covariate at the 10% significance level.

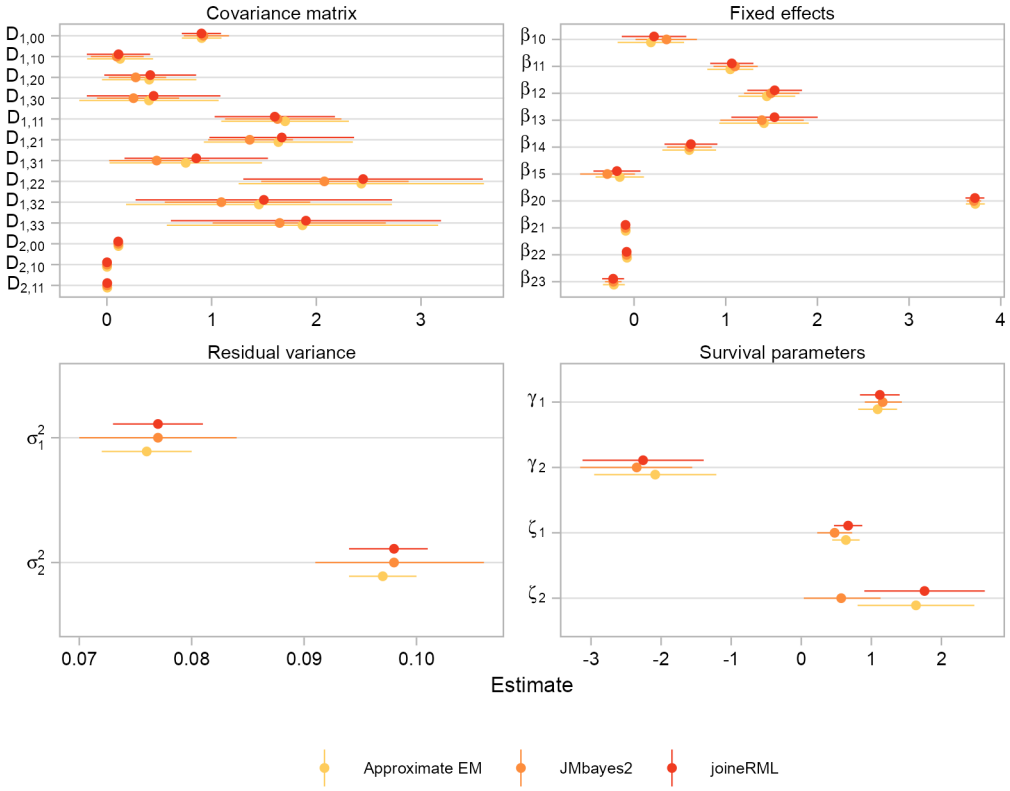


Figure 7.7: Parameter estimates (with 95% confidence/credible) intervals presented for the 'final' joint model on the PBC data (7.4) obtained by the approximate EM algorithm and two existing, established methods `joineRML` and `JMbayes2`. This serves as a graphical companion to the results in Table 7.6. The results are divided into parameter groups for presentation purposes.

For the purpose of the exercise undertaken in this chapter, we are content with this final model (7.4). There is no need for further reduction, since the longitudinal and survival residuals meet typical assumptions, and our parameter estimates align closely with those obtained using existing software. We can now turn our attention towards establishing measures of prognostic performance by implementation of the methods outlined in Section 6.4.

7.6.1 Example dynamic predictions

We focus-in on a few subjects in the PBC data who 'fit' different archetypes of longitudinal follow-up. For the full data, the lower and upper quartiles of follow-up length are three and nine time-points. One profile we want to generate predictions for is failure within the first quartile: Investigating whether the model (7.4) adequately 'dampens' the survival probability $\Pr\left(T_i^* > u_j | T_i^* > t, \mathcal{D}_i(t), \mathcal{D}; \hat{\boldsymbol{\Omega}}\right) \; \forall \; u_j > t$. We also want to investigate such predictions for a profile who survives until the third quantile $t \approx 9$, but fails in the final quantile of follow-up, i.e. they appear to be following the protective trajectory of bilirubin and/or albumin, but fail 'at the final hurdle'. Finally, we investigate these predictions for subjects censored during follow-up.

In each case we set $\mathcal{D}_i(t)$ to be the available data at the last observed longitudinal time t, and obtain the estimates $\hat{\boldsymbol{\pi}}_i = (\hat{\pi}(u_1|t), \ldots, \hat{\pi}(u_f|t))^\top, \; u_j > t \; \forall \; j = 1, \ldots, f$ by the Monte Carlo scheme outlined (on a time-by-time basis) in Section 6.3.3 with 200 simulated draws. The acceptance rate is controlled to be approximately 23% in all cases.

Case study: Almost immediate failure

In the PBC data, the subject `id` 1 dies after approximately 1.1 years of follow-up. They had `histologic` stage 4 (i.e. cirrhosis of the liver) at commencement of the study were older than average (`age` $= 0.83$). They have two observed longitudinal measurements, with final observed $t = 0.53$. Their estimated random effects were $\hat{\boldsymbol{b}}_{i1} = (2.05, 0.73, 1.52, 0.87)^\top$ and $\hat{\boldsymbol{b}}_{i2} = (-0.52, -0.04)^\top$; the directionality of these random effects indicate an increased hazard for the subject ($\hat{\gamma}_1 = 1.090$, $\hat{\gamma}_2 = -2.084$). There were 126 observed mortalities after time t.

In Figure 7.8 we observe that the estimated survival probabilities leading up to the true survival time, $\hat{\pi}_i(u_j|t) \; \forall \; u_j < T_i^*$, are sharply decreasing as $u_j \to T_i^*$, with $\hat{\pi}_i(T_i^*|t) = 0.402$ i.e. given this subject's baseline characteristics and estimates for random effects $\hat{\boldsymbol{b}}_i$ we predict a 60% chance of failure at their actual failure time. Additionally, the predicted probabilities $\hat{\pi}_i(u_j|t) \; \forall \; u_j > T_i^*$ continue in sharp decrease towards zero; within a year of T_i^* we assign a 90% chance of failure.

Case study: Failure in final stages of follow-up

We elect the subject `id` 21, who fails at time $T_i^* = 10.02$, approximately one year after their last recorded longitudinal follow-up time $t = 9.01$. Like subject `id` 1 in the previous case study, this subject has `histologic` stage 4 indicating cirrhosis but is slightly older with `age` $= 1.34$. The subject's estimated random effects $\hat{\boldsymbol{b}}_{i1} = (-1.33, 0.84, 0.91, 0.24)^\top$

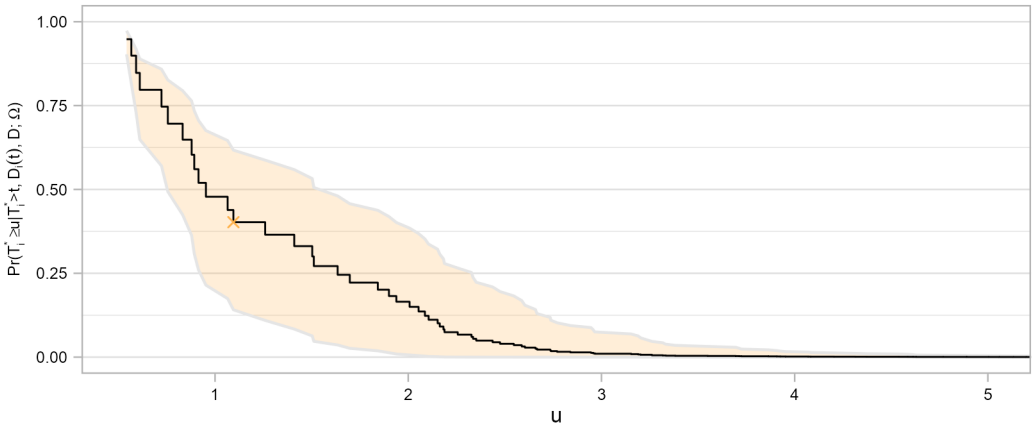


Figure 7.8: Estimated survival probabilities $\hat{\pi}_i(u|t)$ for subject `id` 1 who failed after 1.10 years of follow-up (denoted by the orange 'X'). The 95% confidence interval for the estimate is signified by the shaded orange band. The x-axis is truncated at $u = 5$. Due to `plotmath` restrictions in `R` $\hat{\boldsymbol{\Omega}}$ is displayed as $\boldsymbol{\Omega}$.

and $\hat{\boldsymbol{b}}_{i2} = (0.28, -0.05)^\top$ are protective for both serum bilirubin and albumin at the start of the clinical trial, but have since progressed unfavourably, increasing the hazard.

Figure 7.9 shows the estimates for the probability $\hat{\pi}_i(u|t)$. We see that we obtain the estimate $\hat{\pi}_i(T_i^*|t) = 0.50$; within half a year of further follow-up the point-estimate approximately halves and we obtain $\hat{\pi}_i(10.51|t) = 0.27$. Clearly, after initially protective values, the subject's estimated increased trajectory (given by $\hat{\boldsymbol{b}}_i$) rapidly increases the hazard for $u_j > t$.

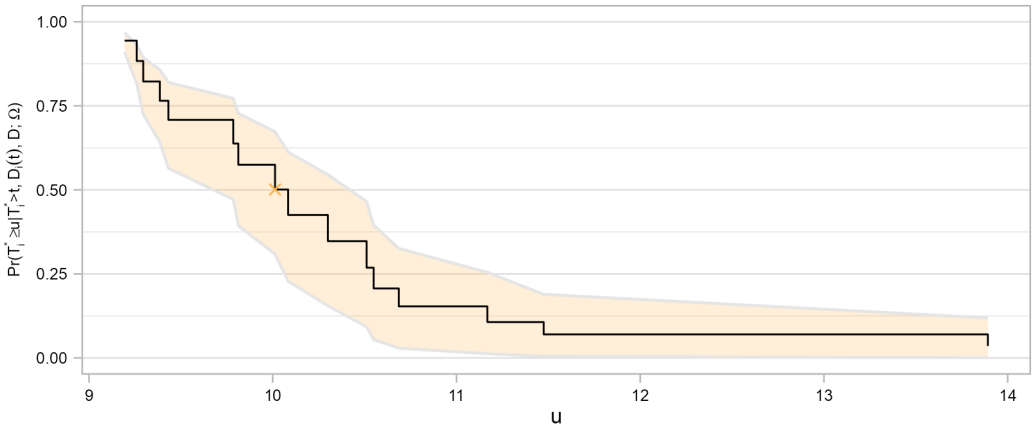


Figure 7.9: Estimated survival probabilities $\hat{\pi}_i(u|t)$ for subject `id` 21 who failed after 10.02 years of follow-up (denoted by the orange 'X'). The 95% confidence interval for the estimate is signified by the shaded orange band. Due to `plotmath` restrictions in `R` $\hat{\boldsymbol{\Omega}}$ is displayed as $\boldsymbol{\Omega}$.

Case study: Censoring midway through follow-up

We now consider those who did not die during follow-up but instead dropped out of the Mayo clinic study for other reasons i.e. were censored. In actuality, as mentioned in Section 7.1, in the PBC data *two* clinically relevant endpoints exist: Transplantation *or* mortality. We therefore want to investigate the predicted survival probabilities of those who drop out during follow-up after surviving for differing periods of time. We elect `id` 5, who undergoes transplantation at $T_5 = 4.12$ years; `id` 7, who drops out of the study at $T_7 = 6.85$; and `id` 312, who drops out at $T_{312} = 3.99$. We are interested in whether the estimated $\hat{\pi}_i(u|t)$ is poor for those who dropped out of the study/were transplanted, since only those with advanced liver disease i.e. poorest disease prognosis undergo transplantation of the liver, or if their leaving the study may not have been related to their disease.

The three plots of $\hat{\pi}_i(u|t)$ in Figure 7.10 illustrate that the prognosis for these individuals greatly differed. Subject `id` 5 who underwent transplantation almost immediately after their last observed follow-up time shares a similar – if narrower – set of predicted survival probabilities to `id` 312, who had their final observed longitudinal follow-up time approximately 1.5 years earlier; both of these subjects have estimated random effects which place them well above (below) average trajectories in bilirubin (albumin) at the start of the study, as well as over its follow-up. Interestingly however, both of these subjects are below the mean `age`, and neither had cirrhosis at their baseline visit; perhaps indicating that these estimated random effects trajectories $\hat{\boldsymbol{b}}_i$ hold the greatest influence on the subject's prognosis. This is apparently cemented by $\hat{\pi}_7(u|t)$: Their estimated $\hat{\boldsymbol{b}}_i$ placing them above (below) the population trajectory for albumin (bilirubin) whilst sharing other patient characteristics with `id` 5. The information carried over time by the biomarkers change prognosis; the longitudinal information is critical to understanding their survival prospects.

7.6.2 ROC and AUC for the final model

It would be exhaustive to carry out, and present, dynamic predictions for all 312 patients in the PBC study; the case studies carried out in the previous section intended to provide an insight into how these may be used in a prognostic predictive setting. Instead, we attempt to estimate measures of discrimination and calibration, as outlined in Sections 6.4.2 and 6.4.3, respectively, for our chosen model (7.4).

We elect three time 'windows' (following the same definition given in Section 6.4.1). We set the first as $w_1 = (2, 3.5]$, which contains approximately 25% of all failure times; the next $w_2 = (3.5, 7]$ containing 30%; and $w_3 = (7, 14]$ with 20%. We don't consider predictions based on the truncated data at $t < 2$. The 34 failures occurring in w_1 happen, on average,

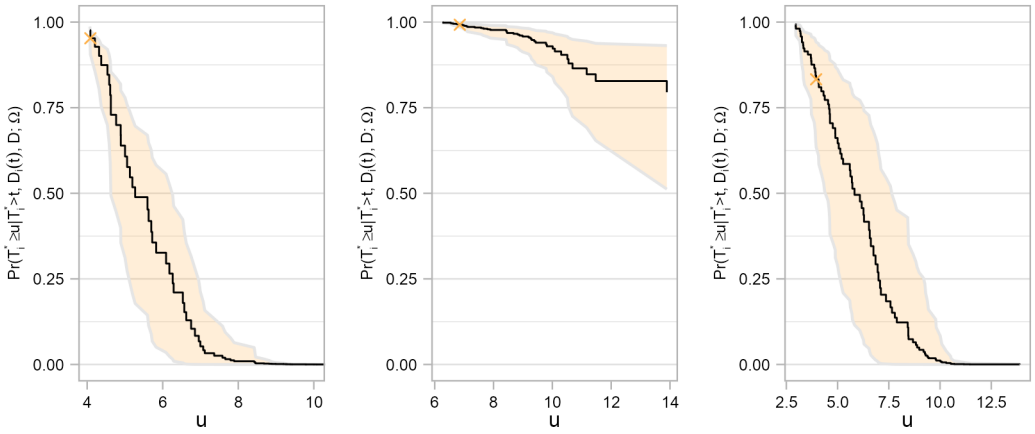


Figure 7.10: Estimated survival probabilities $\hat{\pi}_i(u|t)$ for subject `id`s 5 (left pane), 7 (middle pane) and 312. The censor times were $T_5 = 4.12,\ T_7 = 6.85,\ T_{312} = 3.99$ (denoted by the orange 'X' in each pane). The 95% confidence interval for the estimate is signified by the shaded orange band. Due to `plotmath` restrictions in `R` $\hat{\boldsymbol{\Omega}}$ is displayed as $\boldsymbol{\Omega}$.

once every 0.04 years; in w_2 the 42 failures happen doubly slowly with rate 0.08 years; w_3 contains 31 observed failure times occurring in a more staggered fashion with average time between failures 0.22 years.

We estimate the survival probabilities $\hat{\pi}_i(w)$ by the first-order estimate we introduced in Section 6.3.2, comparing with analogous AUCs for the joint model fit by the 'gold standard' `JMbayes2`, which samples from its MCMC chains to perturb $\hat{\boldsymbol{\Omega}}$ and an M-H scheme to sample from $f\left(\boldsymbol{b}_i|T_i, \Delta_i, \boldsymbol{Y}_i; \hat{\boldsymbol{\Omega}}\right)$. For each $w_1, \ldots, w_3$ we present the ROC curve, along with the AUC from our implementation as well as `JMbayes2`'s.

Following the methodology we outlined in Sections 6.4.2–6.4.4 we obtain the un/corrected AUCs and prediction errors presented in Table 7.7, with ROC curves presented in Figure 7.11. The model (7.4) performs well at distinguishing between those who fail in the elected time windows $w_1, \ldots, w_3$ and those who do not since the median corrected $\text{AUC}_{w_i} > 0.8, i = 1, \ldots, 3$. The predictive errors inform us that in w_1 only 6% of individuals are misclassified, with a larger portion in w_3 ($\approx 13\%$). The predictive error increases commensurately with the decreasing AUC as we move 'later' in follow-up, indicating that both the model's accuracy and discrimination ability deteriorates slightly as follow-up progresses.

For reference (*once again noting and emphasising the differences between underlying models*), `JMbayes2` evaluates the AUC to be 0.884, 0.862, and 0.780 for each time window respectively. We note the `JMbayes2` AUC estimate lies within the confidence intervals for all three windows, and in the central tendency (i.e. interquartile range) for windows w_2 and w_3, indicating some agreement here. The factor of performance drop-off $\text{AUC}_{w_1}/\text{AUC}_{w_3}$

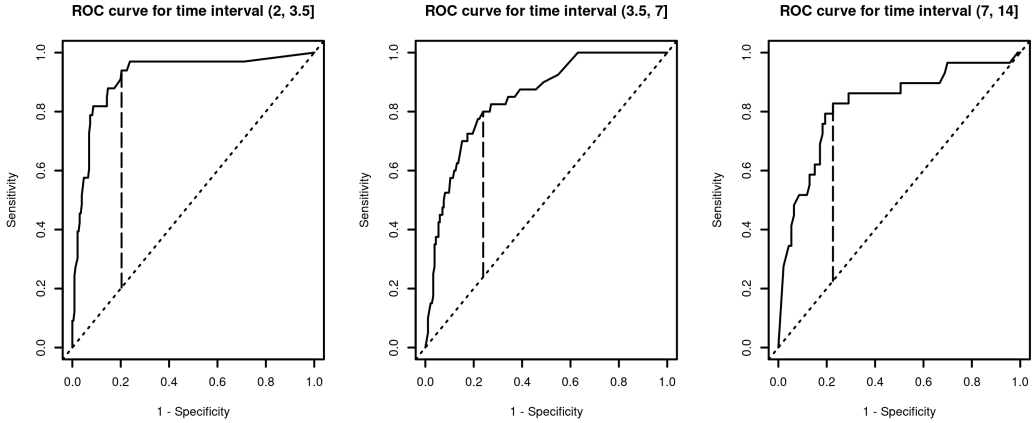


Figure 7.11: ROC curves for the final model evaluated at three rolling time-windows. The apparent (i.e. uncorrected) area under the curve is given in the lower-right hand legend for each window. The dashed vertical line represents the optimal trade-off between the TPR (y-axis) and FPR (x-axis) as identified by the maximal Youden's J statistic (6.13) across probabilistic thresholds.

is approximately equal between approaches, but we note that the AUCs obtained from `JMbayes2` decrease much more sharply between windows; possibly telling of the current-value parameterisation in later stages of follow-up over-predicting the cumulative risk.

	AUC			$\widehat{\text{PE}}$		
	Median	IQR	95% CI	Median	IQR	95% CI
w_1	0.910	[0.895, 0.929]	[0.877, 0.979]	0.065	[0.058, 0.072]	[0.046, 0.086]
w_2	0.838	[0.817, 0.866]	[0.777, 0.912]	0.093	[0.088, 0.102]	[0.075, 0.119]
w_3	0.812	[0.770, 0.834]	[0.716, 0.907]	0.131	[0.119, 0.156]	[0.105, 0.169]

Table 7.7: Corrected estimates for AUC and prediction errors $\widehat{\text{PE}}$ calculated using the methodology outlined in Section 6.4.4.

Attempting now to complete inference on the ROC curves produced by the model fit with approximate EM in Figure 7.11, we turn attention to the maximal Youden indices, J_Y (6.13), for each window investigated.

For w_1 we obtain maximal J_Y at the probability threshold $c_{w_1} = 0.91$, with resultant $\text{TPR}_{w_1} = 0.94$ and $\text{FPR}_{w_1} = 0.20$. Using the probabilistic threshold c_{w_1} on the predicted survival probabilities, $\hat{\pi}_i(w_1) < 0.91$, demonstrates a strong ability to identify individuals who experience mortality within w_1 (indicated by the high TPR). The low FPR observed is largely due to this high threshold: The prognostic model is conservative in labelling instances as negative thereby reducing the risk of incorrectly labelling true negatives as false positives. Four example true/false positives and true/false negatives for this window – with the same idea extending to all windows – are given in Figure 7.12.

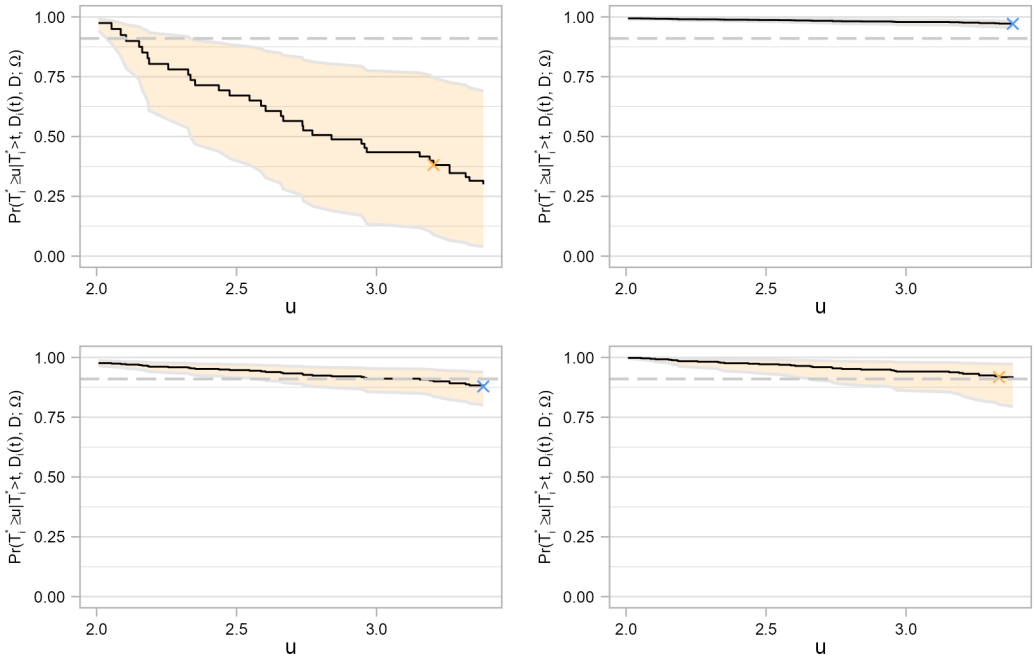


Figure 7.12: Four example $\hat{\pi}_i(u_j|T_{\text{start}})\ \forall\ u_j \in \boldsymbol{u}_{w_1}$. Clockwise from top-left: True positive, true negative, false negative, and false positive. An orange band represents the 95% confidence interval. An orange 'X' displays their true event time $T_{\text{start}} < T_i^* < h$ and blue one $T_i > h$.

Performing the same investigation in w_2 and w_3, we obtain $c_{w_2} = 0.84$ and $c_{w_3} = 0.71$ based on the maximal J_Y. These thresholds produce similar levels of sensitivity $\text{TPR}_{w_2} = 0.80$, $\text{TPR}_{w_3} = 0.83$ whilst returning FPRs as low as we observed in w_1, $\text{FPR}_{w_2} = 0.24$, $\text{FPR}_{w_3} = 0.23$.

We note for the model (7.4) we obtain a consistent trade-off between TPR and FPR: The model performs well over time in terms of its ability to correctly identify those who fail, and provides constant trade-off between true and false positives. The decreasing threshold identified by J_Y in each window implies adaptability in identification of those we deem to require intervention (e.g. $\hat{\pi}_i(w_x) \leq c_{w_x}$) over follow-up, specifically that we become more lenient (i.e. less ready) in declaring failures over time.

Finally, we examine the extra accuracy (6.18) gained by jointly modelling albumin with serum bilirubin. In Table 7.6 although both biomarkers are strongly associated with mortality, one notes $\left|\frac{\hat{\gamma}_1}{\text{SE}(\hat{\gamma}_1)}\right| > \left|\frac{\hat{\gamma}_2}{\text{SE}(\hat{\gamma}_2)}\right|$. We therefore investigate the impact of 'dropping' albumin – i.e. recovering the univariate joint model fit to serum bilirubin in Section 7.4.1 – and establish the improvement to model accuracy gained by its inclusion. We denote the model (7.4) $\mathcal{M}_1$ and the univariate bilirubin joint model $\mathcal{M}_0$. In the same manner that Table 7.7 presents the performance measures for $\mathcal{M}_1$, Table 7.8 presents these measures for $\mathcal{M}_0$. Briefly comparing we do not note any difference in median AUC estimates,

but that the prediction error appears to increase as follow-up progresses more sharply under $\mathcal{M}_0$. That is, the two models are able to offer similar discrimination between non-/failures, but the actual predicted probabilities are systemically higher/lower than the 'true' probabilities under $\mathcal{M}_0$, perhaps indicating that albumin 'tempers' the increased hazard from serum bilirubin.

The presence of albumin in the joint model does not appear to contribute to prediction accuracy at all in w_1, with (median [IQR]) $R(w_1) = -0.023$ $[-0.063,\ 0.007]$. However, its inclusion significantly improves prediction accuracy with $R(w_2) = 0.085$ $[0.063,\ 0.116]$ and $R(w_3) = 0.142$ $[0.110,\ 0.182]$. This appears to indicate that albumin does *not* improve prediction accuracy in the study's infancy, but aspects of its trajectory in the later stages of follow-up account for approximately 6–11% 'extra' accuracy in the 'middle' of follow-up and 11–18% in the last half of follow-up. This gain in accuracy may be useful to practitioners, outweighing the cost of a slightly more complex model.

Visual representation of the AUC, $\widehat{\text{PE}}$ and R measures discussed in this exercise is provided in Figure 7.13 for greater ease of comparison. The measures are found following the methodology in Section 6.4.4 expanded to *two* models per bootstrapped data set.

	AUC			$\widehat{\text{PE}}$		
	Median	IQR	95% CI	Median	IQR	95% CI
w_1	0.912	[0.892, 0.935]	[0.877, 0.979]	0.064	[0.056, 0.071]	[0.046, 0.086]
w_2	0.830	[0.811, 0.853]	[0.771, 0.894]	0.104	[0.096, 0.112]	[0.085, 0.130]
w_3	0.813	[0.788, 0.850]	[0.731, 0.902]	0.152	[0.141, 0.167]	[0.121, 0.188]

Table 7.8: Corrected estimates for AUC and prediction errors $\widehat{\text{PE}}$ calculated using the methodology outlined in Section 6.4.4 for $\mathcal{M}_0$.

7.7 Conclusions

In this chapter, we undertook a comprehensive application to the oft-used primary biliary cirrhosis data set. We began with setting-out our motivations and described the data at hand, with attention placed on available longitudinal biomarkers we believed to be associated with mortality and baseline information. We developed a survival sub-model and meticulously crafted multiple longitudinal GLMMs independently at the outset before undertaking univariate joint modelling to establish association between each biomarker and mortality. This was followed by multivariate joint modelling, where we grouped together biomarkers into broad groups of liver function in order to whittle down to, at first a trivariate joint model, and then in proceeding to our final joint model.

When considering candidate 'final' models in the latter stages of model building, we compared the estimated $\hat{\boldsymbol{\Omega}}$ across existing established software `joineRML` and `JMbayes2` as well

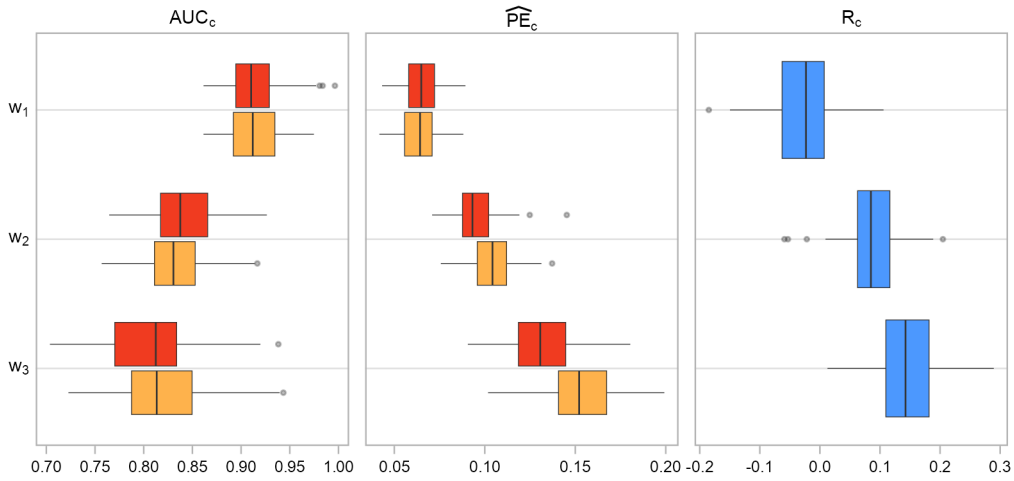


Figure 7.13: Boxplots of corrected AUC, $\widehat{\text{PE}}$ and R for the univariate joint model $\mathcal{M}_0$ (orange), and the bivariate $\mathcal{M}_1$ (red). The subscript c denotes that these are corrected measures. This provides visual comparison of the performance measures provided in Tables 7.7 and 7.8 along with resultant measure of the 'extra' accuracy gained (6.18), wherein the boxplots are blue since they do not belong to *one* model.

as the approximate EM algorithm. We noted good agreement across approaches, with the approximate EM algorithm achieving relatively quicker model convergence.

With our final model (7.4) in hand, we applied post-hoc analyses from Chapter 6 and determined the prognostic accuracy and discriminatory power of our final model. The final model (7.4) appears to be in accordance with the long-standing result of serum bilirubin being considered a strong indicator of liver disease progression (Murtaugh et al., 1994; Rizopoulos, 2012a).

Throughout the joint modelling processes we utilised the methodologies delineated in Chapters 3 and 4, namely through use of the `R` package `gmvjoint` (for more information see Appendix D). When we compared results with those from `JMbayes2`, which operates under the Bayesian paradigm, we alluded to the differences between maximum likelihood and MCMC approaches. Generally, one expects parameter estimates, along with their associated uncertainty, obtained through MCMC methods to surpass those derived from maximum likelihood estimation: The sampling technique employed by MCMC explores a wider range of the parameter space, potentially revealing complex interactions that may go unnoticed by the MLE approach; additionally, the symmetric distribution of errors assumed by MLE may be unrealistic when compared to the true posterior distribution sampled by MCMC. Results in Tables 7.4–7.6 may have agreed (i.e. across MLE and MCMC approaches) to a larger extent by e.g. using a different choice of priors, or analogously defining an unspecified baseline hazard in the Bayesian fit for greater model parity.

Chapter 8

Conclusion & Future Work

In the years since Wulfsohn and Tsiatis (1997) introduced joint modelling, clinical trials in various disease areas have consistently gathered extensive longitudinal biomarker data. This development has introduced both challenges and opportunities. Multivariate data offers enhanced predictive discrimination, but ignoring its multivariate nature means overlooking potentially informative correlations between longitudinal trajectories. However, traditional parameter estimation methods, which involve multidimensional integrals, present substantial computational and statistical challenges.

This aim of this thesis was to explore alternative approaches for faster fitting of joint models for survival and multivariate longitudinal data. By repurposing an underutilized approximate EM algorithm (Bernhardt et al., 2015), we aimed to reduce the computational burden associated with traditional methods of fitting joint models, which successfully lead to faster fitting. Additionally, we moved beyond the commonly restrictive longitudinal specifications used in these models, opting for more flexible and potentially complex specifications.

In Chapter 1 we sought to provide something of a background to joint modelling itself: The data analysis challenges it was borne out of, and early methods it superseded, with particular focus on the seminal article introducing joint modelling as it appears in the thesis (Wulfsohn and Tsiatis, 1997). We then outlined the many ways in which joint models have 'evolved' since this first simplified model. For instance, the longitudinal sub-models featuring more complex random effects specifications and/or the presence of more than one longitudinal response. We also drew attention to literature which eschewed imposition of the Gaussian (i.e. LMM) – Survival (i.e. Cox PH) meta-model. The joint models as they would appear in Chapters 2 and 4 then having precedence in literature.

We then defined multivariate joint models under, what we termed, the 'classic' framework in Chapter 2, introducing the K linear mixed effects models constructing the longitudinal

sub-models as well as the survival sub-model. The observed data likelihood and the form of its constituent densities for such joint models were given, and we outlined semiparametric maximum likelihood via the EM algorithm under a joint modelling framework. With particular focus on the E-step we gave the form of necessary expectations taken against the conditional distribution of the random effects, i.e. the multidimensional integration required for parameter updates, outlining numerical techniques to evaluate such integrals appearing in literature. We then detailed a few techniques to obtain standard errors in the parameter estimates *after* convergence of such an EM algorithm. We finally introduced simulation of the longitudinal response, as well as the survival process before combining these and outlining simulation under a joint modelling framework as a whole.

The EM algorithm outlined in Chapter 2 serves as our main workhorse for parameter inference in the thesis. In Chapter 3 we reaffirm our motivation in undertaking multivariate joint model fits (i.e. potentially many longitudinal responses and/or random effects), but draw attention to the prohibitive computational expense of existing routines to evaluate multidimensional integrals which may limit implementation of such joint models.
To remedy this we introduce an approximate EM algorithm originally proposed in Bernhardt et al. (2015) under a joint model of multivariate longitudinal data with a logistic (in place of the Cox PH model we consider), which allow these intractable multidimensional integrals to be evaluated against a univariate normal distribution.
We outlined starting values, convergence details, as well as the algorithm itself i.e. how one achieves convergence. We outlined the parameter updates under this approximate EM approach, followed by extensive simulation studies which aimed to establish the strengths and potential sensitivities of inference by the approximate EM algorithm. We concluded that the algorithm appears to perform well, save for potential issues with scaling with sample size n, which is not necessarily exclusive to this line of ML estimation. We closed this chapter by comparisons to established software (Hickey et al., 2018a), where we noted very similar parameter estimates which were obtained in a faster manner, particularly for large K, and conducted sensitivity analyses.

Being confident we have an approach under which we can quickly fit multivariate joint models, we began Chapter 4 by restating that the 'classic' Gaussian assumption imposed on *all* longitudinal responses across Chapters 2 and 3 may be inappropriate in clinical settings. We then introduce how one formulates generalised linear mixed models (GLMMs) as well as stating the advantages of such models (e.g. we can eschew transformations on data and consider their 'most natural' distribution) and preordained uptick in computational demand when electing this modelling avenue.
Attention then focused more specifically on GLMMs within joint modelling: Elucidating the need to accommodate diverse longitudinal outcomes and highlighting the exclusively Bayesian paradigm such joint models are fit under in the literature owing to the form

expectations would take if one proceeded by e.g. EM instead, before highlighting the usage of the approximation used in Chapter 3 in circumventing this restrictive formulation. This neatly sets the scene for (what we coin) 'generalised' multivariate joint models to be defined.

We consider five candidate exponential families, surplus to the Gaussian, to model the conditional mean of the longitudinal response: Poisson; Gamma; Binomial; Negative binomial; and generalised Poisson. These were chosen in an attempt to capture differing types of data including binary, skewed continuous and counts, including count data exhibiting over- and/or under-dispersion. The densities of these chosen families are presented, and we detail parameter estimation for these generalised multivariate joint models with attention on the fixed effects $\boldsymbol{\beta}$ and dispersion parameters $\boldsymbol{\sigma}$; the survival sub-model and random effects specification not being altered here.

As was undertaken in Chapter 3, we then carried out simulation studies on these joint models. However, rather than emphasis on data characteristics (with an assumption that the behaviour seen in Chapter 3 will *not* drastically change), focus was instead placed on demonstration of parameter estimation under mixtures of these different families, in trivariate and five-variate simulations. We also presented univariate simulations for the considered families which housed dispersion parameters. We concluded that the non-Gaussian specifications of the longitudinal sub-models did not adversely affect the parameter estimates; the approximate EM algorithm providing an efficient manner to fit these joint models under maximum likelihood. However, the performance of the algorithm was notably weakest when a binary response was considered, and over-dispersed generalised Poisson, with both shortcomings discussed.

We closed the chapter by investigating/adapting a pre-existing heuristic for the number of quadrature nodes one should use, where we noted results were in-line with existing literature for joint models, as well as a brief excursion into potential usage of the approximation in a Monte Carlo EM scheme. Across the presented simulation results presented in Chapters 3 and 4, we identified something of a trade-off between efficient computation and model performance (as determined by e.g. 95% coverage probabilities).

Taking a step back from these joint models which have dominated the thesis up to this point, we next investigated the normal approximation proffered by Bernhardt et al. (2015) and adapted in Chapter 3 in Chapter 5. Here we sought to provide some background to the approximation by drawing attention to similar techniques used in literature previously (Rizopoulos, 2012a) and contrasting with that used in the thesis. We next set out the objectives to be assessed by simulation: Essentially to investigate whether the approximation appears to be reasonable; followed by how we would achieve this. Given a pre-determined set of parameter values, $\boldsymbol{\Omega}^{(\text{TRUE})}$, we sampled from the true posterior $f\left(\boldsymbol{b}_i|\mathcal{D}_i;\boldsymbol{\Omega}^{(\text{TRUE})}\right)$ and compared this against the approximate normal distribution

$N\big(\hat{\boldsymbol{b}}_i, \hat{\Sigma}_i\big)$. We concluded that the approximation utilised throughout appeared reasonable, even with non-Gaussian conditional distributions on the response. However, we additionally observed the variance-covariance estimate $\hat{\Sigma}_i$ overestimated the variance of the true distribution $f\big(\boldsymbol{b}_i|\mathcal{D}_i; \boldsymbol{\Omega}^{(\text{TRUE})}\big)$. We hypothesised this could be remedied by future development of a scale factor applied to $\hat{\Sigma}_i$ (i.e. to be 'closer' to the true normal distribution), or alternative specifications of the random effects altogether.

With details on parameter estimation and obtention of standard errors given in Chapters 2–4, i.e. 'having' the fitted joint model, in Chapter 6 we turn attention to a suite of post-hoc analyses one can carry out on the said model. We laid out calculation of residuals for the longitudinal and survival sub-models and provided examples for each. We then surveyed the readily-available inference joint models enjoy with respect to hypothesis testing of the parameters present in the sub-models as well as comparing two nested joint models. However we noted here issues with comparing non-nested models, and that much of the literature here is underdeveloped. We then proceeded from basic inference on a joint model to obtaining predictions following routines in existing literature (Rizopoulos, 2012a). We outlined estimation of the probabilities of interest, as well as how these can be used to establish time-dependent measures of discrimination and calibration for the fitted joint model, including methods to adequately establish predictive contributions from e.g. additional biomarkers.

Chapter 7 acts as an amalgamation of the multivariate joint models introduced in Chapter 2, the methods for fast and flexible fitting we developed in Chapters 3 and 4, supplemented by post-hoc analyses we outlined in Chapter 6. In this chapter we sought to provide a comprehensive modelling process applied to the motivating set of clinical data, PBC, introduced in Chapter 1. We began removed from the joint modelling instead focusing on the longitudinal and survival sub-models, in each case identifying a model which offered a good trade-off between goodness-of-fit and model simplicity. We then established univariate association before considering three separate multivariate joint models, split by broad categories of liver function. Over the next few stages we then arrived at a 'final' bivariate joint model. With this in tow, we undertook subject-specific dynamic predictions and presented the associated AUC measures as the focal point. Across model-building stages, as well as in these discriminatory measures, we compared the parameter estimates from the joint models fit by the approximate EM algorithm with existing software, noting good agreement, with approximate EM achieving model convergence in a faster manner.

8.1 Future work

8.1.1 Extensions to the survival sub-model

Throughout the work presented, we have solely considered a Cox PH model for the time-to-event process in the joint model. One obvious extension would be considering a *cause-specific* (i.e. competing risks) model. Examples of joint models with a competing risks survival sub-model include Williamson et al. (2008), Li et al. (2010) and Rustand et al. (2023). Such joint models would indeed be more useful in circumstances where patients can experience multiple events of interest: For instance death or disease recurrence; recurrent events such as re-admission to hospital; or a succession of events such as transition between (e.g. worsening of) disease states (Hickey et al., 2018b).

The univariate as well as competing risks survival sub-models are still very much couched within a Cox PH model, an assumption of which being proportional hazards. Although there exist plenty of common techniques to validate this modelling assumption, such a check has not materialised (to the best of the author's knowledge) for joint models. One popular alternative to the Cox PH model are accelerated failure time (AFT) models. Instead of modelling the hazard, AFTs model the time-to-event by either transforming (say taking logarithms of) the survival times, or through imposition of a distributional assumption (e.g. Weibull, Gompertz) with extra parameters to be estimated. In instances where the PH model is unappealing due to the assumption of proportional hazards potentially not being met, or some evidence that the survival times follow a specific distribution, a joint model with an AFT sub-model may then be appealing, with two examples being Tseng et al. (2005), and Khan and Basharat (2022).

Finally, within the specification of the survival sub-model in the joint modelling framework in (2.2), one could consider *separate* γ terms on the random intercepts and slopes. This would provide inference on e.g. risk associated with having a higher than average biomarker measurement at the outset, as well as that associated with quicker than average decline in the biomarker. This specification could be flexibly specified on a per-biomarker basis.

8.1.2 Power for joint models

Throughout the thesis, we referred to consternation surrounding potential lack of power for joint models, with the considerations in Appendix A.3 being set out for a *univariate* joint model. To the best of the author's knowledge, no real steps have been taken in wider literature to investigate power calculations for the multivariate case. Riley et al. (2019) proffer, under a multivariate Cox PH model, that the model should be fitted to data with a certain sample size n and number of events $\sum_{i=1}^{n} \Delta_i$ *relative* to the number of covariates

and define their criterion in terms of 'events per parameter'. Extending this line of enquiry to a *multivariate* joint model, we may infer that fitting, say, a $K = 10$ joint model to a set of data with relatively low sample size and incidence of failure will not have adequate power, and run risk of over-fitting both the estimates $\boldsymbol{\gamma}$ as well as the underlying hazard. Investigations herein were outside the scope of the thesis, and remain a vital avenue for future research if multivariate joint models are to be used in practice.

8.1.3 Further exponential family members

Given the focus on GLMMs in Chapter 4, where we consider only five families when modelling the response (excluding the traditional Gaussian), there is a very clear opportunity to allow for yet further flexibility in the specification of the longitudinal sub-model. For continuous responses, the (multivariate) t_ν-distribution (with degrees of freedom ν to be estimated) could be implemented.

In instances where the response is continuous but bounded $\boldsymbol{Y}_i \in [0, 1]$ e.g. representing a probability/percentage then the Beta family could be used to model the conditional mean of such a response. Something of an extension is ordered Beta regression (Kubinec, 2023) where the interval $[0, 1]$ is cut into ordinal segments with some meaningful transition between them; the linear predictor then modelling the cumulative probability of being in each segment.

Longitudinal responses which are counts in nature also enjoy a great breadth of potential families. For responses where underdispersion is thought to be present, the Gamma-Count distribution (Zeviani et al., 2014) could be employed, for over- and under-dispersed data the Conway-Maxwell-Poisson (CMP) distribution (Conway and Maxwell, 1962; Shmueli et al., 2005), and its reparameterisation via the mean (Huang, 2017) could be used. This CMP distribution is similar to the generalised Poisson, but with no restriction on its dispersion model.

Outside of the families themselves one could implement, we could additionally consider different 'regimes' to each response's sub-model. That is, in addition to the dispersion sub-models we introduced in Section 4.3.2 we could have modelled other parts of the (sub) modelling process. Perhaps the most obvious avenues here are modelling a structural excess of zeroes (zero-inflation, an example being Zhu et al. (2018)); a structural absence of zeroes (zero-truncation); or hurdle models, which essentially combine these approaches by first modelling the probability of zero versus non-zero counts and then modelling the non-zero counts separately, using a truncated count distribution.

One could therefore view the approximate EM algorithm introduced as being somewhat nascent, with the families we do consider in Chapter 4 perhaps providing a basis for a flexible joint modelling approach.

8.1.4 Methods for faster computation

The approximation in Chapter 3 reduces computation time by effectively 'flattening' multidimensional integrals to one-dimensional ones which are appraised by quadrature. This *alone* doesn't lead to the computation times observed in Chapters 3, 4 and 7, for instance the M-step update for the survival parameters we outlined in Section 3.3.5 still presents a computational bottleneck. This lead to the use of `C++` (interfaced from `R` by `Rcpp` (Eddelbuettel and François, 2011) and `RcppArmadillo` (Eddelbuettel and Sanderson, 2014)) in practicality, which greatly reduced the impact of these bottlenecks as mentioned in Section 3.7 with an example given in Appendix D.6.

We noted in Section 3.4.2 that the approximate EM algorithm appeared to slow down fairly harshly with increased sample size n. One promising avenue in literature recently is the use of a linear scan algorithm to greatly improve computational efficiency under semiparametric maximum likelihood via EM (Li et al., 2022) One promising avenue in literature recently is the use of a linear scan algorithm to greatly improve computational efficiency under semiparametric maximum likelihood via EM. A linear scan algorithm sequentially processes data points in a single pass, rather than iterating over them all multiple times. This method significantly reduces the computational burden by avoiding the repetitive calculations typically required by EM. By doing so, it enhances the speed and efficiency of its EM algorithm, potentially making it a valuable tool for handling large datasets. This approach is available in `R` package `FastJM` (Li, 2022). Unfortunately despite the computational promise here, the package is restricted to the univariate Gaussian case, and so wasn't considered in comparisons with other software in Chapters 3 and 7.

The approximate EM algorithm presented in this thesis appears to scale well with many biomarkers, exhibiting a near-linear increase in computation time with $K = 2 \rightarrow 7$, corresponding to an increase in the dimension of the random effects $q = 2 \rightarrow 14$ in Sections 3.4.3, 3.5, and 4.5.5. Despite this, in circumstances where *very* many biomarkers exist in the data, inference can become more complex owing to either the lack of sample size and/or event incidence as we discuss in Appendix A.3; we noted the impact in Sections 3.4.3 and 3.5, furthermore in our application to PBC data in Chapter 7 we opted for several smaller models in lieu of one large model with this consternation in mind.
However, should practitioners wish to simply 'use all' data, these inferential issues would make the methodology in this thesis inadequate. We could then consider following, for example Li and Luo (2017) and Li et al. (2021), in use of functional principal components analysis on the longitudinal processes. This could lead to the development of hybrid approaches which utilise some dimension reduction method in tandem with the approximate EM algorithm to use 'more' data at the outset, whilst still enjoying faster computation.

We provided either analytical or numerical (see Appendix A.1) solutions to the differen-

tiation necessary for statistical inference in Sections 3.3 and 4.4.2. One exciting avenue for further development in joint modelling both generally and in terms of computational efficiency could be interfacing with automatic differentiation (AD) via AD Model Builder (Fournier et al., 2012), implemented in `R` by template model builder (`TMB`, Kristensen et al. (2016)). Without getting into specific details, `TMB` allows the user to near-instantaneously calculate first and second order derivatives for likelihood functions which are either pre-existing, or user-defined (as a joint model would be), therefore greatly decreasing the time taken per EM iteration.

Glossary

This glossary outlines common notation used in more than one chapter of the thesis.

K The number of longitudinal responses to be jointly modelled.

$\boldsymbol{Y}_{ik}$ The k^{th} longitudinal response observed for subject i, $k = 1, \ldots, K$

m_{ik} The number of observations for subject i on the k^{th} response.

m_i The total number of observed measurements for subject i across all K responses i.e. $m_i = \sum_{k=1}^{K} m_{ik}$.

T_i^* The (possibly unobserved) failure time for subject i.

C_i The (possibly unobserved) censoring time for subject i.

T_i The event time for subject i, defined as whichever occurs first out of T_i^* and C_i.

Δ_i Failure indicator, taking value 1 if $T_i = T_i^*$ and zero if $T_i = C_i$.

X_{ik} The design matrix for the fixed effects for subject i for the k^{th} response.

Z_{ik} The design matrix for the random effects for subject i for the k^{th} response.

$\boldsymbol{\beta}_k$ The fixed effects on the k^{th} response. The collection across all K responses is $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_K^\top\right)^\top$.

$\boldsymbol{b}_{ik}$ The random effects for subject i for the k^{th} response. The collection across all K responses is $\boldsymbol{b}_i = \left(\boldsymbol{b}_{i1}{}^\top, \ldots, \boldsymbol{b}_{iK}{}^\top\right)^\top$.

ε_k The residuals for the linear mixed effects model.

q_k The dimension of random effects as specified for the k^{th} longitudinal model.

q The total dimension of random effects across all K responses i.e. $q = \sum_{k=1}^{K} q_k$.

$\lambda_i(t)$ The hazard of instantaneous failure at time t for subject i.

$\lambda_0(t)$ The baseline hazard (unspecified) at time t.

$\boldsymbol{S}_i$ The vector of baseline covariates used in the survival sub-model.

$\boldsymbol{\zeta}$ The vector of coefficients associated with the baseline covariates $\boldsymbol{S}_i$.

γ_k The association parameter 'linking' the k^{th} longitudinal response with the hazard by the shared random effects $\boldsymbol{b}_{ik}$.

$\boldsymbol{W}_k(t)$ Vector function of time at time t corresponding to the random effects structure on the k^{th} random effects.

$\boldsymbol{\Phi}$ The collection of survival parameters $\boldsymbol{\Phi} = \left(\boldsymbol{\gamma}^\top, \boldsymbol{\zeta}^\top\right)^\top$.

$\boldsymbol{\Omega}$ The vector of parameters to be estimated.

L The length of $\boldsymbol{\Omega}$.

$\hat{\boldsymbol{x}}$ The maximum likelihood estimate for some parameter $\boldsymbol{x}$.

$\boldsymbol{x}^{(m)}$ The *current* estimate for some parameter $\boldsymbol{x}$ at iteration m.

$g(\cdot)$ Used to denote 'some function'.

$\mathbb{E}[\cdot]$ The expectation taken 'generally'/'across all'.

$\mathbb{E}_i[\cdot]$ The expectation taken specifically with respect to subject is design measures and estimated random effects.

$\tilde{\mathbb{E}}_i[\cdot]$ The approximate expectation.

$\ell(\cdot)$ The log-likelihood.

$\ell_i(\cdot)$ The log-likelihood contribution from subject i.

$s_i(\cdot)$ The score function or gradient vector.

$\mathrm{H}_i(\cdot)$ The hessian function i.e. generates/returns the matrix of second derivatives.

ϱ The number of quadrature weights and abscissae to use.

$\boldsymbol{w}$ The vector of weights used in quadrature routines $\boldsymbol{w} = (w_1, \ldots, w_\varrho)^\top$.

$\boldsymbol{v}$ The vector of abscissae used in quadrature routines $\boldsymbol{w} = (v_1, \ldots, v_\varrho)^\top$.

$\mathcal{I}$ An information matrix

ν Either the scale of the Gompertz baseline hazard used in simulation, or the degrees of freedom in a t_ν-distribution.

α The rate of the Gompertz baseline hazard used in simulation.

$\overset{\text{appx.}}{\sim}$ Denotes 'approximately distributed by'.

$\hat{\boldsymbol{b}}_i$ The vector which maximises the complete data likelihood for subject i at $\boldsymbol{\Omega}^{(m)}$.

$\hat{\Sigma}_i$ The variance of the estimate $\hat{\boldsymbol{b}}_i$.

ξ_1, ξ_2 Thresholds use to determine convergence on the absolute and relative scales, respectively.

υ Determines which of ξ_1, ξ_2 is used on a given element of $\boldsymbol{\Omega}$.

ω The failure rate in a simulation $\omega = n^{-1} \sum_{i=1}^{n} \Delta_i$.

r The maximal number of observations a subject can have in a simulation, across all K.

$\boldsymbol{\eta}_i$ The linear predictor.

$h(\cdot)$ A link function relating the conditional mean of the response to the linear predictor with known inverse $h^{-1}(\cdot)$.

$\tilde{\boldsymbol{\eta}}_i$ The linear predictor for the dispersion model.

$\boldsymbol{W}_{ik}$ The design matrix for the dispersion parameters for subject i for the k^{th} response.

$\boldsymbol{\sigma}_k$ The dispersion parameter on the k^{th} response.

$\tilde{h}(\cdot)$ A link function relating the dispersion of the response to its linear predictor with known inverse $\tilde{h}^{-1}(\cdot)$.

$\hat{Y}$ A fitted value for the response Y.

$\hat{r}$ An unscaled residual.

$\hat{r}^{(P)}$ A Pearson residual.

$\hat{r}^{(CS)}$ A Cox-Snell residual.

$\mathcal{M}$ A fitted joint model. $\mathcal{M}_0$ and $\mathcal{M}_1$ are used to denote a simpler and more complex model, respectively.

$\pi_i(u|t)$ The probability that subject i survives to future time u given their available information at time t.

$S(\cdot)$ The survival function.

w A window of interest in which a joint model's predictive accuracy is to be ascertained.

T_{start} The starting time of the window.

h The horizon time of the window

$\boldsymbol{u}_w$ The vector of length f containing all failure times in w.

$\hat{\pi}_i(w)$ The estimated probability that subject i survives w.

AUC The area under the ROC curve.

$\widehat{\text{PE}}$ The estimated prediction error.

Appendix A

Extra information and derivations

A.1 Numerical differentiation techniques

In numerous places, we have utilised numerical differentiation to obtain score vectors and hessian matrices necessary to form parameter updates when faced with particularly unattractive log-likelihood functions. For instance in updating the dispersion parameters $\boldsymbol{\sigma}$ in the Negative binomial GLMM (Section 4.4.3), or differentiating the conditional expectation of the survival log-density in Section 3.3.5. We note these are not strictly speaking the *scores* (instead the *gradient*), but we continue with this statistical faux pas regardless. The analytical implementations would simply be tedious and time-consuming in such cases. We therefore outline forward and central differencing to obtain the Score vector for P-vector parameter $\boldsymbol{\beta}$ in Algorithms A.1 and A.2 respectively, and a three-point central difference formula used to obtain the Hessian matrix in A.3.

Note that the algorithm presented to obtain the Hessian matrix is overkill for any case when the parameter vector of interest is of length one. In this scenario, we simply find the second derivative by a three-point central differencing routine given as the sub-routine on line 16 of Algorithm A.3.

Algorithm A.1 Forward differencing for score calculation

Require:

ℓ: The log-likelihood function, returns scalar value.

$\boldsymbol{\beta}$: Parameter vector we want to differentiate log-likelihood with respect to.

h: Step size for differencing, typically cube root of machine precision.

Ensure:

$\boldsymbol{S}$: Score vector

$\ell_0 \leftarrow \ell(\boldsymbol{\beta})$

$\boldsymbol{S} \leftarrow \mathbf{0}$

for $j \leftarrow 1$ to P **do** ▷ Loop over P parameters

  $\boldsymbol{\beta}^* \leftarrow \boldsymbol{\beta}$

  $\beta_j^* \leftarrow \beta_j + h$

  $\ell^* \leftarrow \ell(\boldsymbol{\beta}^*)$

  $S_j \leftarrow (\ell^* - \ell_0)/h$

end for

return $\boldsymbol{S}$

Algorithm A.2 Central differencing for score calculation

Require:

ℓ: The log-likelihood function, returns scalar value.

$\boldsymbol{\beta}$: Parameter vector we want to differentiate log-likelihood with respect to.

h: Step size for differencing, typically cube root of machine precision.

Ensure:

$\boldsymbol{S}$: Score vector

$\boldsymbol{S} \leftarrow \mathbf{0}$

for $j \leftarrow 1$ to P **do** ▷ Loop over P parameters

  $\boldsymbol{\beta}^+ \leftarrow \boldsymbol{\beta};\ \boldsymbol{\beta}^- \leftarrow \boldsymbol{\beta}$

  $\beta_j^+ \leftarrow \beta_j + h;\ \beta_j^- \leftarrow \beta_j - h$

  $\ell^+ \leftarrow \ell(\boldsymbol{\beta}^+);\ \ell^- \leftarrow \ell(\boldsymbol{\beta}^-)$

  $S_j \leftarrow (\ell^+ - \ell^-)/2h$

end for

return $\boldsymbol{S}$

Algorithm A.3 Three-point central differencing for Hessian calculation

Require:
ℓ: The log-likelihood function, returns scalar value.
$\boldsymbol{\beta}$: Parameter vector we want to differentiate log-likelihood with respect to.
h: Step size for differencing, typically fourth root of machine precision.
Ensure:
H: Hessian Matrix.

$\mathrm{H} \leftarrow O_{P,P}$ ▷ $P \times P$ matrix of zeroes
$\mathrm{M} \leftarrow \mathbb{I}_P \times h$ ▷ $P \times P$ matrix of zeroes with h on diagonal
if $P > 1$ **then**
  for $j \leftarrow 1$ to $(P-1)$ **do**
    $\boldsymbol{h}_j \leftarrow \mathrm{M}_{:,j}$
    $\mathrm{H}_{j,j} \leftarrow (\ell(\boldsymbol{\beta} + \boldsymbol{h}_j) + \ell(\boldsymbol{\beta} - \boldsymbol{h}_j) - 2 \times \ell(\boldsymbol{\beta}))/h^2$
    for $k \leftarrow (j+1)$ to P **do**
      $\boldsymbol{h}_k \leftarrow \mathrm{M}_{:,k}$
      $\mathrm{H}_{j,k} \leftarrow (\ell(\boldsymbol{\beta}+\boldsymbol{h}_j+\boldsymbol{h}_k) - \ell(\boldsymbol{\beta}+\boldsymbol{h}_j-\boldsymbol{h}_k) - \ell(\boldsymbol{\beta}-\boldsymbol{h}_j+\boldsymbol{h}_k) + \ell(\boldsymbol{\beta}-\boldsymbol{h}_j-\boldsymbol{h}_k))/(4\times h^2)$
      $\mathrm{H}_{k,j} \leftarrow \mathrm{H}_{j,k}$ ▷ Symmetrise
    end for
  end for
  $\boldsymbol{h} \leftarrow \mathrm{M}_{:,P}$
  $\mathrm{H}_{P,P} \leftarrow (\ell(\boldsymbol{\beta} + \boldsymbol{h}) + \ell(\boldsymbol{\beta} - \boldsymbol{h}) - 2 \times \ell(\boldsymbol{\beta}))/h^2$
else
  $\mathrm{H}_{1,1} \leftarrow (\ell(\beta + h) + \ell(\beta - h) - 2 \times \ell(\beta))/h^2$ ▷ NB: scalar argument
end if
return H

A.2 Simulation of failure times for Weibull and Exponential baseline hazards

In Section 2.5.3 we considered only simulation of the Gompertz distribution of event times. For completeness' sake, we show the routine taken for both the exponential and Weibull event times from our re-written Cox PH model (2.32) in turn.

Exponentially distributed event times

Under the exponential distribution, the baseline hazard is determined by the scale parameter $\lambda_0(t) = \nu, \nu > 0$. We then form

$$\begin{aligned}\Lambda(t|\cdot) &= \int_0^t \lambda(u|\cdot)\, du = \int_0^t \nu \exp\{P + Qu\}\, du, \\ &= \nu \exp\{P\} \int_0^t \exp\{Qu\}\, du, \\ &= \nu \exp\{P\} \left[\frac{\exp\{Qu\}}{Q}\right]_0^t = \frac{\nu \exp\{P\}}{Q}(\exp\{Qt\} - 1)\,.\end{aligned}$$

Then, setting the survival function $S(t) = \exp\{-\Lambda(t|\cdot)\}$ equal to a random uniform draw, $U \sim \text{Unif}(0, 1)$, we obtain the simulated survival time T

$$\begin{aligned}\log U = -\Lambda(t|\cdot) &= -\frac{\nu \exp\{P\}}{Q}(\exp\{Qt\} - 1)\,, \\ \exp\{Qt\} &= 1 - \frac{Q \log U}{\nu \exp\{P\}},\end{aligned}$$

and finally,

$$T = \frac{1}{Q} \log\left[1 - \frac{Q \log U}{\nu \exp\{P\}}\right]. \tag{A.1}$$

Weibull distributed event times

Under the Weibull distribution, the baseline hazard is determined by the scale ν and shape α, $\lambda_0(t) = \nu\alpha t^{\alpha-1},\ \alpha, \nu > 0$. The cumulative hazard is then

$$\begin{aligned}\Lambda(t|\cdot) &= \int_0^t \lambda(u|\cdot)\, du = \int_0^t \exp\{P + Qu\}\, \nu\alpha u^{\alpha-1} du, \\ &= \nu\alpha \exp\{P\} \int_0^t \exp\{Qu\}\, u^{\alpha-1} du,\end{aligned}$$

by substitution we then obtain (Austin, 2012)

$$\begin{aligned}\Lambda(t|\cdot) &= \nu\alpha \exp\{P\} \left[\frac{Q \exp\{Qu^{1+\alpha}\}}{1+\alpha}\right]_0^t, \\ &= \frac{\nu\alpha \exp\{P\}\, Q}{1+\alpha}\big(\exp\{Qt^{1+\alpha}\} - 1\big)\,.\end{aligned}$$

Again, setting the survival function equal to a random uniform draw, $U \sim \text{Unif}(0,1)$, we work toward obtention of the simulated survival time T

$$\begin{aligned} \log U &= -\Lambda(t|\cdot)\,, \\ \implies \exp\{Qt^{1+\alpha}\} - 1 &= -\frac{(1+\alpha)\log U}{Q\nu\alpha\exp\{P\}}, \end{aligned}$$

which we can rearrange to obtain our simulated event time,

$$T = \left[\frac{1}{Q}\log\left(1 - \frac{(1+\alpha)\log U}{Q\nu\alpha\exp\{P\}}\right)\right]^{\frac{1}{1+\alpha}}. \tag{A.2}$$

A.3 A note on statistical power for joint models

A *univariate* joint model simulated under a random intercept-and-slope with known variance-covariance D and association parameter γ has its power $\tilde{\beta} = (1-\beta)$ at the one-sided significance level $\tilde{\alpha}$ governed by (Chen et al., 2011)

$$\Xi = \frac{\left(Z_{1-\tilde{\alpha}} + Z_{\tilde{\beta}}\right)}{\gamma^2 G}, \tag{A.3}$$

which informs the rate of failures $H = \Xi/n$. The quantity in the denominator G is controlled by the median failure time T_{Med}, average follow-up time $\bar{t}$ as well as the variance-covariance matrix D with variances D_{11}, D_{22} and known covariance $\text{D}_{21} = \text{D}_{12}$

$$\begin{aligned} G &= \text{D}_{11} + \frac{\text{D}_{22}}{H}\left[\frac{2}{h^2} - e^{-h\bar{t}}\left(\bar{t}^2 + \frac{2\bar{t}}{h} + \frac{2}{h^2}\right)\right] + \frac{2\text{D}_{21}}{H}\left[\frac{1}{h} - e^{-h\bar{t}}\left(\bar{t} + \frac{1}{h}\right)\right], \\ h &= \frac{\log 2}{T_{\text{Med}}}. \end{aligned} \tag{A.4}$$

From the relation between (A.3) and (A.4) we can infer the interplay between our simulation parameters, the simulated data itself, and the power with which one can subsequently estimate γ. Firstly, a larger chosen value for the association parameter itself $|\gamma|$ would result in a smaller necessary sample size n for adequate statistical power; and larger variances in D lead to smaller required sample sizes, with positive covariance D_{21} sharing this same directionality and D_{21} instead demanding larger n.

A.4 Alternative method for calculation of $s_i(\textbf{vech}(\text{D}))$

In 3.2.5, we showed one method for calculation of the scores of (the half vectorisation of) the variance-covariance matrix D in (3.6). This notably involved calculation of the

quantity $\frac{\partial \mathrm{D}}{\partial \boldsymbol{\Omega}_{\mathrm{D}}}$.

This partial derivative takes the form of a cube in computation routines, with each 'slice' the derivative with respect to each element of vech(D) calculated against the score function in (3.6). We can circumvent derivation of this quantity by simply differentiating the conditional expectation of the log-likelihood (2.5) with respect to the (known symmetric) matrix D, thereby obtaining its matrix derivative. Then, taking half-vectorisation of and doubling the off-diagonal contributions (correctly weighting their appearance in both the upper- and lower-triangles of this derivative), we obtain the same score as given in (3.6).

The derivative of the conditional expectation on (2.5) with respect to D is given by

$$\frac{\partial \mathbb{E}_i[\log f(\boldsymbol{b}_i|\mathrm{D})]}{\partial \mathrm{D}} = \left(\frac{1}{2}\mathrm{D}^{-1}\left[\hat{\Sigma}_i + \hat{\boldsymbol{b}}_i\hat{\boldsymbol{b}}_i^\top\right]\mathrm{D}^{-1} - \frac{1}{2}\mathrm{D}^{-1}\right)^\top. \tag{A.5}$$

A.5 Gradient vector of complete data log-likelihood wrt $\boldsymbol{b}_i$

The numerical optimiser used in calculation of $\hat{\boldsymbol{b}}_i$ (3.2) benefits greatly from a gradient function being supplied, else numerical differentiation routines are instead used internally by `optim`, at some computational expense. With this in mind, we demonstrate formation of the gradient vector taken with respect to $\boldsymbol{b}_i$ of the complete data log-likelihood,

$$\frac{\partial \log f\big(\boldsymbol{b}_i, T_i, \Delta_i, \boldsymbol{Y}_i; \hat{\boldsymbol{\Omega}}\big)}{\partial \boldsymbol{b}_i} = \frac{\partial \log f(\boldsymbol{Y}_i|\boldsymbol{b}_i; \hat{\boldsymbol{\Omega}})}{\partial \boldsymbol{b}_i} + \frac{\partial \log f(\boldsymbol{b}_i|\hat{\boldsymbol{\Omega}})}{\partial \boldsymbol{b}_i} + \frac{\partial \log f(T_i, \Delta_i|\boldsymbol{b}_i; \hat{\boldsymbol{\Omega}})}{\partial \boldsymbol{b}_i}. \tag{A.6}$$

We refer to Section 4.4.2 for the case when $\boldsymbol{Y}_i|\boldsymbol{b}_i$ is non-Gaussian, since $\frac{\partial \log f(\boldsymbol{Y}_i|\boldsymbol{b}_i;\hat{\boldsymbol{\Omega}})}{\partial \boldsymbol{b}_i} = \mathrm{Z}_i^\top \dot{\boldsymbol{\eta}}_i$, with the approximated expected value in the definition of $\dot{\boldsymbol{\eta}}_i$ (4.14) corresponding complete data log-likelihood components with no expectation operator on it. For the Gaussian case (not surveyed in Section 4.4.2) this quantity is given by $\frac{\partial \log f(\boldsymbol{Y}_i|\boldsymbol{b}_i;\hat{\boldsymbol{\Omega}})}{\partial \boldsymbol{b}_i} = \mathrm{Z}_i^\top \mathrm{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\eta}_i)$; the derivative of the log density of random effects is trivially given by $\frac{\partial \log f(\boldsymbol{b}_i|\hat{\boldsymbol{\Omega}})}{\partial \boldsymbol{b}_i} = -\mathrm{D}^{-1}\boldsymbol{b}_i$, and finally for the (rewritten) survival log-density (3.11)

$$\begin{aligned}\frac{\partial \log f(T_i, \Delta_i|\boldsymbol{b}_i; \hat{\boldsymbol{\Omega}})}{\partial \boldsymbol{b}_i} = \Delta_i\left[\boldsymbol{F}_i^\top \boldsymbol{\gamma}^*\right] \\ - \boldsymbol{\gamma}^*\left(\mathrm{F}_{\boldsymbol{u}_i}^\top\left[\lambda_0(\boldsymbol{u}_i) \odot \exp\left\{\mathrm{S}_i^\top \boldsymbol{\zeta} + \mathrm{F}_{\boldsymbol{u}_i}(\boldsymbol{b}_i \odot \boldsymbol{\gamma}^*)\right\}\right]\right),\end{aligned} \tag{A.7}$$

where $\odot$ denotes element-wise multiplication; $\mathrm{F}_{\boldsymbol{u}_i}$ is the horizontal concatenation of $\mathrm{F}_{\boldsymbol{u}_i 1}, \ldots, \mathrm{F}_{\boldsymbol{u}_i K}$; and $\boldsymbol{\gamma}^* = (\gamma_1, \gamma_1, \ldots, \gamma_1, \gamma_2, \ldots, \gamma_{K-1}, \gamma_K, \ldots, \gamma_K)^\top$ i.e. the vector with q_k replicates of $\gamma_k \ \forall \ k = 1, \ldots, K$.

A.6 Proof that the GP1 reduces to the Poisson when $\boldsymbol{\varphi}_i = 0$

The log-likelihood of the Poisson is given in (4.6), and the GP-1 in (4.9). Taking $\boldsymbol{\sigma} = 0 \implies \boldsymbol{\varphi}_i = 0$,

$$\begin{aligned}\log f(\boldsymbol{Y}_i|\boldsymbol{b}_i;\boldsymbol{\beta},\boldsymbol{\varphi}_i=0) &= \mathbf{1}^\top \log \boldsymbol{\mu}_i + (\boldsymbol{Y}_i - \mathbf{1})^\top \log \boldsymbol{\mu}_i - \boldsymbol{Y}_i^\top \cancelto{0}{\log \mathbf{1}} \\ &\qquad - \mathbf{1}^\top \log \boldsymbol{Y}_i! - \mathbf{1}^\top \boldsymbol{\mu}_i \\ &= \cancel{\mathbf{1}^\top \log \boldsymbol{\mu}_i} + \boldsymbol{Y}_i^\top \log \boldsymbol{\mu}_i - \cancel{\mathbf{1}^\top \log \boldsymbol{\mu}_i} - \mathbf{1}^\top \boldsymbol{\mu}_i - \mathbf{1}^\top \log \boldsymbol{Y}_i! \\ &= \boldsymbol{Y}_i^\top \log \boldsymbol{\mu}_i - \mathbf{1}^\top \boldsymbol{\mu}_i - \mathbf{1}^\top \log \boldsymbol{Y}_i! \qquad \square\end{aligned} \tag{A.8}$$

A.7 Determining whether a point lies within an ellipse

Suppose we want to determine whether a 'test point' coordinate (x, y) lies within the ellipse with covariance matrix Σ and center point $\boldsymbol{O} = (a, b)$. We calculate the eigenvalues λ_1 and λ_2 and eigenvectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ of the covariance matrix, and use these to calculate the semi-minor and semi-major axes, r_y and r_x, respectively, of the ellipse

$$r_x = \sqrt{\lambda_1 \chi^2_{\alpha,2}}, \qquad r_y = \sqrt{\lambda_2 \chi^2_{\alpha,2}}, \tag{A.9}$$

where $\chi^2_{\alpha,2}$ denotes the quantile of the χ^2_2 distribution on two degrees of freedom which corresponds to probability level α. We then calculate the ellipse's orientation ϑ given as the arctangent between the x axis and $\boldsymbol{O}$. The region bounded by the ellipse is given by

$$\frac{\left(\cos\vartheta(x-a) + \sin\vartheta(y-b)\right)^2}{r_x^2} + \frac{\left(\sin\vartheta(x-a) - \cos\vartheta(y-b)\right)^2}{r_y^2} \leq 1. \tag{A.10}$$

We test the coordinate (x, y) by checking whether or not it satisfies the inequality (A.10). If satisfied, then (x, y) lies *within* the ellipse.

Appendix B

Supplementary Tables

Here, additional tabulated results from Chapters 3, 4, and 7 are presented.

All tabulations for simulation studies carried out in Chapters 3 and 4 have the following summaries reported, where $\{\hat{\boldsymbol{\Omega}}_i\}$ is the set of N joint model estimates and $\boldsymbol{\Omega}^{(\text{TRUE})}$ the true parameter values:

$$
\begin{aligned}
\text{“Mean”} &= \frac{\sum_{i=1}^{N} \hat{\boldsymbol{\Omega}}_i}{N}; \qquad & \text{“SD”} &= \frac{\sum_{i=1}^{N} \left(\hat{\boldsymbol{\Omega}}_i - \text{Mean}\right)}{N-1}; \\
\text{“SE”} &= \frac{\sum_{i=1}^{N} \text{SE}\big(\hat{\boldsymbol{\Omega}}_i\big)}{N}; & & \\
\text{“Bias”} &= \frac{\sum_{i=1}^{N} \left(\hat{\boldsymbol{\Omega}}_i - \boldsymbol{\Omega}^{(\text{TRUE})}\right)}{N}; & & \\
\text{“MSE”} &= \frac{\sum_{i=1}^{N} \left(\boldsymbol{\Omega}^{(\text{TRUE})} - \hat{\boldsymbol{\Omega}}_i\right)^2}{N}; & & \\
\text{“CP”} &= \frac{\sum_{i=1}^{N} I\big(\hat{\boldsymbol{\Omega}}_i - 1.96\text{SE}(\hat{\boldsymbol{\Omega}}_i) < \boldsymbol{\Omega}^{(\text{TRUE})}, \hat{\boldsymbol{\Omega}}_i + 1.96\text{SE}(\hat{\boldsymbol{\Omega}}_i) > \boldsymbol{\Omega}^{(\text{TRUE})}\big)}{N}; & &
\end{aligned}
\tag{B.1}
$$

‘MSE’: Mean square error; ‘CP’: Coverage probability (at 95% level). The measures in (B.1) were chosen as they provide insight into average estimation as well as bias exhibited in the results, both on a point-estimate basis as well as the average squared distance between the estimated and true values. Finally, the coverage probability tells us the proportion of model fits wherein the 95% confidence interval contains the true value. Here the indicator function takes two conditions as input, $I(x, y)$, returning unity if and only if *both* x and y are satisfied.

B.1 Additional results from Chapter 3

B.1.1 Sample size n

Additional results from simulation study undertaken in Section 3.4.2.

	$n = 100$					$n = 250$					$n = 500$					$n = 1000$				
Parameter	Mean (SD)	SE	Bias	MSE	CP	Mean (SD)	SE	Bias	MSE	CP	Mean (SD)	SE	Bias	MSE	CP	Mean (SD)	SE	Bias	MSE	CP
$D_{1,1} = 0.250$	0.244 (0.041)	0.063	-0.006	0.002	0.99	0.247 (0.027)	0.032	-0.003	0.001	0.99	0.249 (0.021)	0.021	-0.001	0.000	0.98	0.252 (0.015)	0.015	0.002	0.000	0.96
$D_{3,1} = 0.125$	0.128 (0.034)	0.048	0.003	0.001	0.99	0.127 (0.021)	0.025	0.002	0.000	0.96	0.124 (0.016)	0.016	-0.001	0.000	0.96	0.126 (0.011)	0.011	0.001	0.000	0.95
$D_{5,1} = 0.125$	0.125 (0.033)	0.047	0.000	0.001	1.00	0.126 (0.023)	0.024	0.001	0.001	0.99	0.125 (0.016)	0.016	0.000	0.000	0.93	0.125 (0.011)	0.011	0.000	0.000	0.96
$D_{2,2} = 0.090$	0.090 (0.015)	0.023	0.000	0.000	0.97	0.090 (0.010)	0.011	0.000	0.000	0.96	0.090 (0.007)	0.007	0.000	0.000	0.97	0.090 (0.005)	0.005	0.000	0.000	0.94
$D_{3,3} = 0.250$	0.248 (0.048)	0.063	-0.002	0.002	0.99	0.251 (0.030)	0.032	0.001	0.001	0.96	0.247 (0.022)	0.021	-0.003	0.001	0.96	0.253 (0.017)	0.015	0.003	0.000	0.93
$D_{5,3} = 0.125$	0.124 (0.035)	0.048	-0.001	0.001	0.99	0.127 (0.021)	0.025	0.002	0.000	0.98	0.124 (0.017)	0.016	-0.001	0.000	0.92	0.125 (0.011)	0.011	0.000	0.000	0.93
$D_{4,4} = 0.090$	0.090 (0.015)	0.022	0.000	0.000	0.99	0.089 (0.010)	0.011	-0.001	0.000	0.93	0.089 (0.007)	0.007	-0.001	0.000	0.95	0.089 (0.005)	0.005	-0.001	0.000	0.94
$D_{5,5} = 0.250$	0.244 (0.045)	0.063	-0.006	0.002	0.99	0.253 (0.031)	0.033	0.003	0.001	0.98	0.250 (0.021)	0.022	0.000	0.000	0.96	0.249 (0.014)	0.015	-0.001	0.000	0.95
$D_{6,6} = 0.090$	0.087 (0.014)	0.022	-0.003	0.000	0.99	0.090 (0.011)	0.011	0.000	0.000	0.92	0.090 (0.006)	0.007	0.000	0.000	0.96	0.090 (0.005)	0.005	0.000	0.000	0.96
$\beta_{10} = 2.000$	2.000 (0.078)	0.103	0.000	0.006	0.98	1.992 (0.055)	0.055	-0.008	0.003	0.93	1.995 (0.035)	0.037	-0.005	0.001	0.94	1.999 (0.025)	0.026	-0.001	0.001	0.96
$\beta_{11} = -0.100$	-0.102 (0.037)	0.045	-0.002	0.001	0.97	-0.106 (0.023)	0.023	-0.006	0.001	0.93	-0.105 (0.016)	0.016	-0.005	0.000	0.95	-0.105 (0.011)	0.011	-0.005	0.000	0.94
$\beta_{12} = 0.100$	0.100 (0.053)	0.076	0.000	0.003	1.00	0.099 (0.036)	0.039	-0.001	0.001	0.94	0.100 (0.027)	0.026	0.000	0.001	0.95	0.098 (0.018)	0.018	-0.002	0.000	0.94
$\beta_{13} = -0.200$	-0.210 (0.126)	0.148	-0.010	0.016	0.97	-0.196 (0.066)	0.078	0.004	0.004	0.98	-0.192 (0.054)	0.052	0.008	0.003	0.94	-0.199 (0.036)	0.036	0.001	0.001	0.95
$\beta_{20} = -2.000$	-1.989 (0.078)	0.107	0.011	0.006	0.97	-2.007 (0.056)	0.056	-0.007	0.003	0.95	-2.000 (0.036)	0.037	0.000	0.001	0.97	-2.000 (0.026)	0.026	0.000	0.001	0.94
$\beta_{21} = 0.100$	0.099 (0.036)	0.045	-0.001	0.001	0.96	0.105 (0.021)	0.023	0.005	0.000	0.96	0.106 (0.014)	0.016	0.006	0.000	0.97	0.106 (0.010)	0.011	0.006	0.000	0.93
$\beta_{22} = -0.100$	-0.099 (0.052)	0.079	0.001	0.003	0.99	-0.102 (0.034)	0.040	-0.002	0.001	0.96	-0.099 (0.024)	0.026	0.001	0.001	0.98	-0.100 (0.016)	0.018	0.000	0.000	0.98
$\beta_{23} = 0.200$	0.182 (0.120)	0.151	-0.018	0.015	0.98	0.208 (0.076)	0.079	0.008	0.006	0.95	0.202 (0.052)	0.052	0.002	0.003	0.96	0.194 (0.037)	0.036	-0.006	0.001	0.95
$\beta_{30} = 2.000$	2.008 (0.078)	0.105	0.008	0.006	0.99	1.991 (0.049)	0.056	-0.009	0.003	0.98	1.998 (0.037)	0.037	-0.002	0.001	0.93	2.004 (0.025)	0.026	0.004	0.001	0.95
$\beta_{31} = -0.100$	-0.106 (0.032)	0.044	-0.006	0.001	0.97	-0.101 (0.022)	0.023	-0.001	0.000	0.95	-0.102 (0.014)	0.016	-0.002	0.000	0.97	-0.105 (0.010)	0.011	-0.005	0.000	0.94
$\beta_{32} = 0.100$	0.095 (0.056)	0.078	-0.005	0.003	0.98	0.100 (0.037)	0.040	0.000	0.001	0.97	0.098 (0.025)	0.026	-0.002	0.001	0.97	0.101 (0.018)	0.018	0.001	0.000	0.95
$\beta_{33} = -0.200$	-0.209 (0.100)	0.150	-0.009	0.010	1.00	-0.192 (0.059)	0.079	0.008	0.004	0.99	-0.198 (0.053)	0.052	0.002	0.003	0.94	-0.204 (0.037)	0.036	-0.004	0.001	0.97
$\sigma_1^2 = 0.160$	0.161 (0.009)	0.012	0.001	0.000	0.99	0.159 (0.005)	0.006	-0.001	0.000	0.96	0.160 (0.004)	0.004	0.000	0.000	0.98	0.160 (0.003)	0.003	0.000	0.000	0.98
$\sigma_2^2 = 0.160$	0.160 (0.008)	0.012	0.000	0.000	1.00	0.159 (0.005)	0.006	-0.001	0.000	0.99	0.160 (0.004)	0.004	0.000	0.000	0.99	0.160 (0.003)	0.003	0.000	0.000	0.96
$\sigma_3^2 = 0.160$	0.159 (0.010)	0.012	-0.001	0.000	0.95	0.161 (0.006)	0.006	0.001	0.000	0.96	0.160 (0.004)	0.004	0.000	0.000	0.96	0.161 (0.003)	0.003	0.001	0.000	0.97
$\gamma_1 = 0.500$	0.548 (0.242)	0.297	0.048	0.060	1.00	0.494 (0.131)	0.142	-0.006	0.017	0.97	0.518 (0.096)	0.091	0.018	0.009	0.95	0.501 (0.067)	0.061	0.001	0.004	0.91
$\gamma_2 = -0.500$	-0.573 (0.243)	0.298	-0.073	0.064	0.99	-0.492 (0.137)	0.140	0.008	0.019	0.95	-0.509 (0.083)	0.089	-0.009	0.007	0.96	-0.499 (0.067)	0.060	0.001	0.004	0.91
$\gamma_3 = 0.500$	0.527 (0.231)	0.305	0.027	0.053	0.98	0.509 (0.140)	0.138	0.009	0.019	0.93	0.493 (0.102)	0.091	-0.007	0.010	0.90	0.497 (0.069)	0.061	-0.003	0.005	0.92
$\zeta = -0.200$	-0.249 (0.445)	0.520	-0.049	0.199	0.96	-0.192 (0.272)	0.268	0.008	0.073	0.96	-0.197 (0.183)	0.180	0.003	0.033	0.97	-0.178 (0.107)	0.123	0.022	0.012	0.96

Table B.1: Parameter estimates for differing sample sizes $n \in \{100, 250, 500, 1000\}$. 'Mean (SD)' denotes the average estimated value with standard deviation of the parameter estimate. 'SE' denotes the mean standard error calculated at each model fit. Coverage probabilities are calculated from $\hat{\boldsymbol{\Omega}} \pm 1.96\text{SE}(\hat{\boldsymbol{\Omega}})$ Median [IQR] elapsed time taken for approximate EM algorithm to converge and standard error calculation for $n = 100$ was 1.799 [1.452, 2.479] seconds and total computation time was 3.015 [2.651, 3.761] seconds; for $n = 250$ was 3.553 [3.314, 3.769] seconds total computation time was 5.348 [5.075, 5.724] seconds; for $n = 500$ was 8.525 [8.243, 8.886] seconds and total computation time was 11.395 [11.047, 11.976] seconds; and for $n = 1000$ was 26.962 [26.193, 27.591] seconds and total computation time was 32.907 [31.979, 33.530] seconds.

B.1.2 Number of longitudinal responses K

In Section 3.4.3 we investigated the impact on computation time of increasing the number of longitudinal responses K (thereby the dimension of random effects $q = 2K$), with choices of $K \in \{1, 2, 3, 5, 7\}$. Here we present additional results from this simulation study.

Parameter	Mean (SD)	SE	Bias	MSE	CP
$D_{1,1} = 0.250$	0.246 (0.021)	0.020	-0.004	0.000	0.90
$D_{2,2} = 0.060$	0.060 (0.005)	0.005	0.000	0.000	0.96
$\beta_{10} = 2.000$	1.998 (0.037)	0.036	-0.002	0.001	0.92
$\beta_{11} = -0.100$	-0.104 (0.012)	0.013	-0.004	0.000	0.95
$\beta_{12} = 0.100$	0.093 (0.025)	0.025	-0.007	0.001	0.93
$\beta_{13} = -0.200$	-0.197 (0.053)	0.051	0.003	0.003	0.93
$\sigma_1^2 = 0.160$	0.161 (0.004)	0.004	0.001	0.000	0.93
$\gamma_1 = 0.500$	0.499 (0.093)	0.093	-0.001	0.009	0.96
$\zeta = -0.200$	-0.223 (0.174)	0.160	-0.023	0.030	0.95

Table B.2: Parameter estimates for $K = 1$. Median [IQR] elapsed time taken for approximate EM algorithm to converge and standard error calculation was 2.173 [2.103, 2.232] seconds and total computation time was 3.510 [3.423, 3.579] seconds.

Parameter	Mean (SD)	SE	Bias	MSE	CP
$D_{1,1} = 0.250$	0.246 (0.019)	0.021	-0.004	0.000	0.99
$D_{2,2} = 0.060$	0.060 (0.005)	0.005	0.000	0.000	0.96
$D_{3,3} = 0.250$	0.251 (0.021)	0.021	0.001	0.000	0.98
$D_{4,4} = 0.060$	0.059 (0.004)	0.005	-0.001	0.000	0.96
$\beta_{10} = 2.000$	1.995 (0.031)	0.036	-0.005	0.001	0.96
$\beta_{11} = -0.100$	-0.106 (0.012)	0.013	-0.006	0.000	0.95
$\beta_{12} = 0.100$	0.100 (0.024)	0.025	0.000	0.001	0.96
$\beta_{13} = -0.200$	-0.198 (0.045)	0.051	0.002	0.002	0.96
$\beta_{20} = -2.000$	-1.993 (0.035)	0.036	0.007	0.001	0.97
$\beta_{21} = 0.100$	0.105 (0.013)	0.013	0.005	0.000	0.94
$\beta_{22} = -0.100$	-0.100 (0.025)	0.026	0.000	0.001	0.96
$\beta_{23} = 0.200$	0.192 (0.047)	0.052	-0.008	0.002	0.97
$\sigma_1^2 = 0.160$	0.160 (0.004)	0.004	0.000	0.000	0.94
$\sigma_2^2 = 0.160$	0.159 (0.004)	0.004	-0.001	0.000	0.93
$\gamma_1 = 0.500$	0.513 (0.096)	0.096	0.013	0.009	0.95
$\gamma_2 = -0.500$	-0.504 (0.099)	0.095	-0.004	0.010	0.93
$\zeta = -0.200$	-0.193 (0.155)	0.163	0.007	0.024	0.99

Table B.3: Parameter estimates for $K = 2$. Median [IQR] elapsed time taken for approximate EM algorithm to converge and standard error calculation was 3.920 [3.776, 4.034] seconds and total computation time was 6.236 [6.117, 6.330] seconds.

Parameter	Mean (SD)	SE	Bias	MSE	CP
$D_{1,1} = 0.250$	0.247 (0.020)	0.021	-0.003	0.000	0.96
$D_{2,2} = 0.060$	0.059 (0.005)	0.005	-0.001	0.000	0.96
$D_{3,3} = 0.250$	0.251 (0.020)	0.022	0.001	0.000	0.96
$D_{4,4} = 0.060$	0.060 (0.005)	0.005	0.000	0.000	0.96
$D_{5,5} = 0.250$	0.245 (0.019)	0.021	-0.005	0.000	0.96
$D_{6,6} = 0.060$	0.060 (0.005)	0.005	0.000	0.000	0.92
$\beta_{10} = 2.000$	1.998 (0.036)	0.037	-0.002	0.001	0.93
$\beta_{11} = -0.100$	-0.104 (0.014)	0.013	-0.004	0.000	0.93
$\beta_{12} = 0.100$	0.101 (0.024)	0.026	0.001	0.001	0.96
$\beta_{13} = -0.200$	-0.201 (0.051)	0.052	-0.001	0.003	0.97
$\beta_{20} = -2.000$	-1.994 (0.036)	0.037	0.006	0.001	0.97
$\beta_{21} = 0.100$	0.104 (0.015)	0.014	0.004	0.000	0.91
$\beta_{22} = -0.100$	-0.095 (0.028)	0.026	0.005	0.001	0.92
$\beta_{23} = 0.200$	0.200 (0.050)	0.052	0.000	0.002	0.98
$\beta_{30} = 2.000$	1.996 (0.036)	0.037	-0.004	0.001	0.97
$\beta_{31} = -0.100$	-0.105 (0.013)	0.014	-0.005	0.000	0.94
$\beta_{32} = 0.100$	0.102 (0.024)	0.026	0.002	0.001	0.96
$\beta_{33} = -0.200$	-0.201 (0.052)	0.052	-0.001	0.003	0.96
$\sigma^2_1 = 0.160$	0.160 (0.004)	0.004	0.000	0.000	0.97
$\sigma^2_2 = 0.160$	0.160 (0.004)	0.004	0.000	0.000	0.98
$\sigma^2_3 = 0.160$	0.161 (0.004)	0.004	0.001	0.000	0.99
$\gamma_1 = 0.500$	0.527 (0.105)	0.098	0.027	0.012	0.93
$\gamma_2 = -0.500$	-0.500 (0.093)	0.098	0.000	0.009	0.98
$\gamma_3 = 0.500$	0.508 (0.106)	0.098	0.008	0.011	0.92
$\zeta = -0.200$	-0.199 (0.146)	0.166	0.001	0.021	0.97

Table B.4: Parameter estimates for $K = 3$. Median [IQR] elapsed time taken for approximate EM algorithm to converge and standard error calculation was 6.379 [6.212, 6.654] seconds and total computation time was 9.576 [9.387, 9.845] seconds.

Parameter	Mean (SD)	SE	Bias	MSE	CP
$D_{1,1} = 0.250$	0.243 (0.017)	0.022	-0.007	0.000	0.97
$D_{2,2} = 0.060$	0.059 (0.005)	0.006	-0.001	0.000	0.96
$D_{3,3} = 0.250$	0.246 (0.018)	0.023	-0.004	0.000	0.96
$D_{4,4} = 0.060$	0.059 (0.005)	0.006	-0.001	0.000	0.96
$D_{5,5} = 0.250$	0.244 (0.021)	0.023	-0.006	0.000	0.95
$D_{6,6} = 0.060$	0.060 (0.005)	0.006	0.000	0.000	0.99
$D_{7,7} = 0.250$	0.247 (0.022)	0.023	-0.003	0.001	0.95
$D_{8,8} = 0.060$	0.059 (0.005)	0.006	-0.001	0.000	0.97
$D_{9,9} = 0.250$	0.247 (0.021)	0.023	-0.003	0.000	0.94
$D_{10,10} = 0.060$	0.059 (0.005)	0.006	-0.001	0.000	0.98
$\beta_{10} = 2.000$	1.997 (0.034)	0.039	-0.003	0.001	0.96
$\beta_{11} = -0.100$	-0.106 (0.012)	0.014	-0.006	0.000	0.99
$\beta_{12} = 0.100$	0.100 (0.024)	0.027	0.000	0.001	0.98
$\beta_{13} = -0.200$	-0.201 (0.042)	0.055	-0.001	0.002	0.99
$\beta_{20} = -2.000$	-2.000 (0.031)	0.039	0.000	0.001	0.99
$\beta_{21} = 0.100$	0.105 (0.014)	0.015	0.005	0.000	0.94
$\beta_{22} = -0.100$	-0.103 (0.021)	0.028	-0.003	0.000	1.00
$\beta_{23} = 0.200$	0.200 (0.041)	0.055	0.000	0.002	0.98
$\beta_{30} = 2.000$	1.997 (0.036)	0.039	-0.003	0.001	0.93
$\beta_{31} = -0.100$	-0.103 (0.013)	0.015	-0.003	0.000	0.98
$\beta_{32} = 0.100$	0.103 (0.023)	0.028	0.003	0.001	0.99
$\beta_{33} = -0.200$	-0.197 (0.051)	0.055	0.003	0.003	0.97
$\beta_{40} = -2.000$	-2.004 (0.033)	0.039	-0.004	0.001	0.99
$\beta_{41} = 0.100$	0.102 (0.012)	0.014	0.002	0.000	0.97
$\beta_{42} = -0.100$	-0.097 (0.026)	0.028	0.003	0.001	0.97
$\beta_{43} = 0.200$	0.209 (0.049)	0.055	0.009	0.002	0.97
$\beta_{50} = 2.000$	1.996 (0.035)	0.039	-0.004	0.001	0.97
$\beta_{51} = -0.100$	-0.106 (0.013)	0.015	-0.006	0.000	0.94
$\beta_{52} = 0.100$	0.097 (0.027)	0.028	-0.003	0.001	0.98
$\beta_{53} = -0.200$	-0.200 (0.048)	0.055	0.000	0.002	0.97
$\sigma^2_1 = 0.160$	0.160 (0.004)	0.005	0.000	0.000	0.97
$\sigma^2_2 = 0.160$	0.161 (0.004)	0.005	0.001	0.000	0.97
$\sigma^2_3 = 0.160$	0.160 (0.004)	0.005	0.000	0.000	0.98
$\sigma^2_4 = 0.160$	0.160 (0.004)	0.005	0.000	0.000	0.97
$\sigma^2_5 = 0.160$	0.160 (0.004)	0.005	0.000	0.000	0.97
$\gamma_1 = 0.500$	0.512 (0.115)	0.106	0.012	0.013	0.93
$\gamma_2 = -0.500$	-0.506 (0.088)	0.107	-0.006	0.008	0.99
$\gamma_3 = 0.500$	0.515 (0.097)	0.106	0.015	0.009	0.94
$\gamma_4 = -0.500$	-0.527 (0.095)	0.107	-0.027	0.010	0.96
$\gamma_5 = 0.500$	0.495 (0.095)	0.108	-0.005	0.009	0.98
$\zeta = -0.200$	-0.220 (0.161)	0.174	-0.020	0.026	0.98

Table B.5: Parameter estimates for $K = 5$. Median [IQR] elapsed time taken for approximate EM algorithm to converge and standard error calculation was 16.480 [16.227, 16.766] seconds and total computation time was 21.389 [21.129, 21.713] seconds.

Parameter	Mean (SD)	SE	Bias	MSE	CP
$D_{1,1} = 0.250$	0.246 (0.018)	0.025	-0.004	0.000	0.99
$D_{2,2} = 0.060$	0.059 (0.005)	0.006	-0.001	0.000	0.98
$D_{3,3} = 0.250$	0.242 (0.020)	0.025	-0.008	0.000	0.97
$D_{4,4} = 0.060$	0.059 (0.005)	0.006	-0.001	0.000	0.95
$D_{5,5} = 0.250$	0.251 (0.021)	0.025	0.001	0.000	1.00
$D_{6,6} = 0.060$	0.060 (0.005)	0.006	0.000	0.000	0.96
$D_{7,7} = 0.250$	0.245 (0.022)	0.025	-0.005	0.000	0.95
$D_{8,8} = 0.060$	0.059 (0.005)	0.006	-0.001	0.000	1.00
$D_{9,9} = 0.250$	0.247 (0.022)	0.025	-0.003	0.000	0.95
$D_{10,10} = 0.060$	0.059 (0.005)	0.006	-0.001	0.000	1.00
$D_{11,11} = 0.250$	0.248 (0.022)	0.025	-0.002	0.000	0.96
$D_{12,12} = 0.060$	0.059 (0.005)	0.006	-0.001	0.000	0.99
$D_{13,13} = 0.250$	0.248 (0.020)	0.025	-0.002	0.000	0.97
$D_{14,14} = 0.060$	0.060 (0.006)	0.006	0.000	0.000	0.98
$\beta_{10} = 2.000$	1.996 (0.038)	0.042	-0.004	0.001	0.99
$\beta_{11} = -0.100$	-0.105 (0.014)	0.016	-0.005	0.000	0.97
$\beta_{12} = 0.100$	0.099 (0.026)	0.030	-0.001	0.001	0.98
$\beta_{13} = -0.200$	-0.199 (0.052)	0.060	0.001	0.003	0.99
$\beta_{20} = -2.000$	-2.001 (0.042)	0.042	-0.001	0.002	0.95
$\beta_{21} = 0.100$	0.103 (0.012)	0.016	0.003	0.000	0.99
$\beta_{22} = -0.100$	-0.099 (0.024)	0.030	0.001	0.001	0.98
$\beta_{23} = 0.200$	0.209 (0.057)	0.060	0.009	0.003	0.95
$\beta_{30} = 2.000$	2.004 (0.039)	0.043	0.004	0.002	0.96
$\beta_{31} = -0.100$	-0.104 (0.013)	0.016	-0.004	0.000	0.98
$\beta_{32} = 0.100$	0.097 (0.024)	0.031	-0.003	0.001	1.00
$\beta_{33} = -0.200$	-0.210 (0.054)	0.061	-0.010	0.003	0.96
$\beta_{40} = -2.000$	-1.992 (0.037)	0.043	0.008	0.001	0.95
$\beta_{41} = 0.100$	0.103 (0.012)	0.016	0.003	0.000	0.99
$\beta_{42} = -0.100$	-0.098 (0.025)	0.030	0.002	0.001	0.98
$\beta_{43} = 0.200$	0.196 (0.055)	0.060	-0.004	0.003	0.98
$\beta_{50} = 2.000$	1.992 (0.032)	0.042	-0.008	0.001	0.99
$\beta_{51} = -0.100$	-0.103 (0.012)	0.016	-0.003	0.000	0.99
$\beta_{52} = 0.100$	0.098 (0.025)	0.030	-0.002	0.001	0.98
$\beta_{53} = -0.200$	-0.197 (0.047)	0.060	0.003	0.002	1.00
$\beta_{60} = -2.000$	-1.988 (0.035)	0.043	0.012	0.001	0.98
$\beta_{61} = 0.100$	0.105 (0.013)	0.016	0.005	0.000	0.97
$\beta_{62} = -0.100$	-0.097 (0.026)	0.030	0.003	0.001	0.95
$\beta_{63} = 0.200$	0.196 (0.046)	0.060	-0.004	0.002	1.00
$\beta_{70} = 2.000$	2.002 (0.035)	0.043	0.002	0.001	0.99
$\beta_{71} = -0.100$	-0.104 (0.014)	0.016	-0.004	0.000	0.97
$\beta_{72} = 0.100$	0.101 (0.026)	0.030	0.001	0.001	0.98
$\beta_{73} = -0.200$	-0.205 (0.050)	0.060	-0.005	0.003	0.99
$\sigma_1^2 = 0.160$	0.160 (0.004)	0.005	0.000	0.000	0.97
$\sigma_2^2 = 0.160$	0.160 (0.004)	0.005	0.000	0.000	0.97
$\sigma_3^2 = 0.160$	0.161 (0.004)	0.005	0.001	0.000	1.00
$\sigma_4^2 = 0.160$	0.160 (0.005)	0.005	0.000	0.000	0.98
$\sigma_5^2 = 0.160$	0.160 (0.005)	0.005	0.000	0.000	0.99
$\sigma_6^2 = 0.160$	0.160 (0.004)	0.005	0.000	0.000	0.98
$\sigma_7^2 = 0.160$	0.160 (0.004)	0.005	0.000	0.000	0.99
$\gamma_1 = 0.500$	0.510 (0.107)	0.119	0.010	0.011	0.99
$\gamma_2 = -0.500$	-0.518 (0.103)	0.122	-0.018	0.011	0.98
$\gamma_3 = 0.500$	0.523 (0.093)	0.119	0.023	0.009	0.98
$\gamma_4 = -0.500$	-0.521 (0.095)	0.120	-0.021	0.009	0.99
$\gamma_5 = 0.500$	0.506 (0.095)	0.121	0.006	0.009	1.00
$\gamma_6 = -0.500$	-0.516 (0.101)	0.119	-0.016	0.010	0.98
$\gamma_7 = 0.500$	0.520 (0.100)	0.120	0.020	0.010	0.98
$\zeta = -0.200$	-0.227 (0.146)	0.191	-0.027	0.022	0.98

Table B.6: Parameter estimates for $K = 7$. Median [IQR] elapsed time taken for approximate EM algorithm to converge and standard error calculation was 29.409 [28.669, 34.127] seconds and total computation time was 36.081 [35.352, 40.809] seconds.

B.1.3 Length of follow-up period r

Additional results from simulation study undertaken in Section 3.4.4. Biases are presented in Figure C.1

	$r = 3$					$r = 5$					$r = 10$					$r = 15$				
Parameter	Mean (SD)	SE	Bias	MSE	CP	Mean (SD)	SE	Bias	MSE	CP	Mean (SD)	SE	Bias	MSE	CP	Mean (SD)	SE	Bias	MSE	CP
$D_{1,1} = 0.250$	0.255 (0.025)	0.029	0.005	0.001	0.99	0.252 (0.022)	0.025	0.002	0.000	0.98	0.247 (0.020)	0.021	-0.003	0.000	0.95	0.250 (0.019)	0.020	0.000	0.000	0.97
$D_{3,1} = 0.125$	0.125 (0.019)	0.020	0.000	0.000	0.97	0.122 (0.016)	0.018	-0.003	0.000	0.97	0.124 (0.014)	0.016	-0.001	0.000	0.97	0.125 (0.014)	0.015	0.000	0.000	0.94
$D_{5,1} = 0.125$	0.123 (0.019)	0.020	-0.002	0.000	0.98	0.121 (0.017)	0.018	-0.004	0.000	0.96	0.124 (0.015)	0.016	-0.001	0.000	0.97	0.123 (0.015)	0.016	-0.002	0.000	0.94
$D_{2,2} = 0.090$	0.090 (0.009)	0.009	0.000	0.000	0.93	0.090 (0.008)	0.008	0.000	0.000	0.94	0.091 (0.006)	0.008	0.001	0.000	0.99	0.089 (0.007)	0.007	-0.001	0.000	0.94
$D_{3,3} = 0.250$	0.252 (0.024)	0.029	0.002	0.001	0.97	0.250 (0.022)	0.025	0.000	0.000	0.97	0.248 (0.019)	0.021	-0.002	0.000	0.98	0.246 (0.017)	0.020	-0.004	0.000	0.99
$D_{5,3} = 0.125$	0.123 (0.017)	0.020	-0.002	0.000	0.97	0.123 (0.017)	0.018	-0.002	0.000	0.94	0.125 (0.016)	0.016	0.000	0.000	0.95	0.122 (0.013)	0.015	-0.003	0.000	0.95
$D_{4,4} = 0.090$	0.089 (0.007)	0.009	-0.001	0.000	0.98	0.089 (0.008)	0.008	-0.001	0.000	0.96	0.090 (0.007)	0.007	0.000	0.000	0.94	0.090 (0.007)	0.007	0.000	0.000	0.94
$D_{5,5} = 0.250$	0.249 (0.028)	0.029	-0.001	0.001	0.95	0.249 (0.024)	0.025	-0.001	0.001	0.94	0.250 (0.021)	0.021	0.000	0.000	0.94	0.248 (0.020)	0.020	-0.002	0.000	0.93
$D_{6,6} = 0.090$	0.089 (0.008)	0.009	-0.001	0.000	0.96	0.089 (0.007)	0.008	-0.001	0.000	0.98	0.088 (0.006)	0.007	-0.002	0.000	0.98	0.090 (0.006)	0.007	0.000	0.000	0.97
$\beta_{10} = 2.000$	1.996 (0.035)	0.041	-0.004	0.001	0.99	2.000 (0.039)	0.039	0.000	0.002	0.96	1.995 (0.035)	0.037	-0.005	0.001	0.96	1.998 (0.030)	0.036	-0.002	0.001	0.98
$\beta_{11} = -0.100$	-0.108 (0.015)	0.018	-0.008	0.000	0.96	-0.106 (0.016)	0.017	-0.006	0.000	0.93	-0.106 (0.015)	0.016	-0.006	0.000	0.97	-0.108 (0.016)	0.016	-0.008	0.000	0.92
$\beta_{12} = 0.100$	0.104 (0.028)	0.029	0.004	0.001	0.94	0.103 (0.025)	0.028	0.003	0.001	1.00	0.103 (0.022)	0.026	0.003	0.000	0.98	0.096 (0.026)	0.026	-0.004	0.001	0.92
$\beta_{13} = -0.200$	-0.203 (0.055)	0.058	-0.003	0.003	0.99	-0.201 (0.057)	0.056	-0.001	0.003	0.93	-0.201 (0.048)	0.052	-0.001	0.002	0.98	-0.197 (0.047)	0.051	0.003	0.002	0.94
$\beta_{20} = -2.000$	-2.001 (0.037)	0.041	-0.001	0.001	0.96	-2.001 (0.036)	0.040	-0.001	0.001	0.97	-2.001 (0.035)	0.037	-0.001	0.001	0.97	-2.006 (0.038)	0.036	-0.006	0.001	0.95
$\beta_{21} = 0.100$	0.108 (0.015)	0.018	0.008	0.000	0.97	0.108 (0.014)	0.017	0.008	0.000	0.98	0.107 (0.015)	0.016	0.007	0.000	0.94	0.104 (0.015)	0.016	0.004	0.000	0.94
$\beta_{22} = -0.100$	-0.101 (0.033)	0.029	-0.001	0.001	0.90	-0.097 (0.023)	0.028	0.003	0.001	1.00	-0.100 (0.023)	0.027	0.000	0.001	0.98	-0.101 (0.025)	0.026	-0.001	0.001	0.94
$\beta_{23} = 0.200$	0.203 (0.057)	0.058	0.003	0.003	0.94	0.205 (0.052)	0.056	0.005	0.003	0.97	0.201 (0.051)	0.053	0.001	0.003	0.97	0.207 (0.047)	0.051	0.007	0.002	0.96
$\beta_{30} = 2.000$	1.998 (0.045)	0.041	-0.002	0.002	0.95	2.006 (0.037)	0.039	0.006	0.001	0.96	1.998 (0.035)	0.037	-0.002	0.001	0.96	1.995 (0.035)	0.036	-0.005	0.001	0.94
$\beta_{31} = -0.100$	-0.108 (0.015)	0.018	-0.008	0.000	0.95	-0.107 (0.016)	0.017	-0.007	0.000	0.95	-0.107 (0.014)	0.016	-0.007	0.000	0.97	-0.105 (0.013)	0.016	-0.005	0.000	0.97
$\beta_{32} = 0.100$	0.099 (0.030)	0.029	-0.001	0.001	0.94	0.102 (0.028)	0.028	0.002	0.001	0.98	0.102 (0.024)	0.026	0.002	0.001	0.98	0.100 (0.026)	0.025	0.000	0.001	0.95
$\beta_{33} = -0.200$	-0.196 (0.066)	0.058	0.004	0.004	0.91	-0.202 (0.055)	0.055	-0.002	0.003	0.92	-0.200 (0.044)	0.053	0.000	0.002	0.99	-0.192 (0.047)	0.051	0.008	0.002	0.95
$\sigma_1^2 = 0.160$	0.155 (0.012)	0.013	-0.005	0.000	0.93	0.159 (0.006)	0.007	-0.001	0.000	0.95	0.161 (0.004)	0.004	0.001	0.000	0.95	0.160 (0.003)	0.003	0.000	0.000	0.98
$\sigma_2^2 = 0.160$	0.156 (0.013)	0.013	-0.004	0.000	0.93	0.160 (0.007)	0.007	0.000	0.000	0.96	0.159 (0.004)	0.004	-0.001	0.000	0.96	0.161 (0.003)	0.003	0.001	0.000	0.98
$\sigma_3^2 = 0.160$	0.158 (0.013)	0.014	-0.002	0.000	0.97	0.160 (0.006)	0.007	0.000	0.000	0.97	0.160 (0.004)	0.004	0.000	0.000	0.97	0.160 (0.003)	0.003	0.000	0.000	0.96
$\gamma_1 = 0.500$	0.469 (0.110)	0.106	-0.031	0.013	0.92	0.493 (0.093)	0.099	-0.007	0.009	0.97	0.497 (0.090)	0.095	-0.003	0.008	0.96	0.511 (0.096)	0.092	0.011	0.009	0.93
$\gamma_2 = -0.500$	-0.470 (0.120)	0.106	0.030	0.015	0.93	-0.490 (0.097)	0.097	0.010	0.009	0.96	-0.513 (0.091)	0.092	-0.013	0.008	0.94	-0.525 (0.093)	0.091	-0.025	0.009	0.96
$\gamma_3 = 0.500$	0.477 (0.100)	0.107	-0.023	0.010	0.96	0.502 (0.087)	0.098	0.002	0.007	0.99	0.524 (0.104)	0.094	0.024	0.011	0.95	0.492 (0.074)	0.092	-0.008	0.005	0.99
$\zeta = -0.200$	-0.217 (0.199)	0.178	-0.017	0.039	0.91	-0.206 (0.162)	0.178	-0.006	0.026	0.95	-0.216 (0.174)	0.178	-0.016	0.030	0.96	-0.189 (0.136)	0.177	0.011	0.018	0.98

Table B.7: Parameter estimates for differing sample sizes $r \in \{3, 5, 10, 15\}$. 'Mean (SD)' denotes the average estimated value with standard deviation of the parameter estimate. 'SE' denotes the mean standard error calculated at each model fit. Coverage probabilities are calculated from $\hat{\boldsymbol{\Omega}} \pm 1.96\text{SE}(\hat{\boldsymbol{\Omega}})$ Median [IQR] elapsed time taken for approximate EM algorithm to converge and standard error calculation for $r = 3$ was 12.584 [11.943, 13.470] seconds and total computation time was 15.581 [14.800, 16.344] seconds; for $r = 5$ was 10.529 [10.094, 11.115] seconds and total computation time was 13.539 [13.143, 14.148] seconds; for $r = 10$ was 8.453 [8.124, 8.862] seconds and total computation time was 12.160 [11.743, 12.636] seconds; for $r = 15$ was 7.788 [7.491, 8.043] seconds and total computation time was 12.404 [12.038, 12.720] seconds.

B.1.4 Failure rate ω

Additional results from simulation study undertaken in Section 3.4.5.

	$\omega = 10\%$					$\omega = 30\%$					$\omega = 50\%$				
Parameter	Mean (SD)	SE	Bias	MSE	CP	Mean (SD)	SE	Bias	MSE	CP	Mean (SD)	SE	Bias	MSE	CP
$D_{1,1} = 0.250$	0.249 (0.021)	0.021	-0.001	0.000	0.96	0.248 (0.023)	0.021	-0.002	0.001	0.92	0.249 (0.021)	0.022	-0.001	0.000	0.95
$D_{3,1} = 0.125$	0.125 (0.015)	0.016	0.000	0.000	0.98	0.125 (0.016)	0.016	0.000	0.000	0.94	0.125 (0.015)	0.017	0.000	0.000	0.98
$D_{5,1} = 0.125$	0.125 (0.016)	0.016	0.000	0.000	0.96	0.124 (0.017)	0.016	-0.001	0.000	0.93	0.123 (0.016)	0.016	-0.002	0.000	0.96
$D_{2,2} = 0.090$	0.089 (0.005)	0.007	-0.001	0.000	0.98	0.090 (0.007)	0.007	0.000	0.000	0.94	0.090 (0.008)	0.008	0.000	0.000	0.96
$D_{3,3} = 0.250$	0.248 (0.019)	0.021	-0.002	0.000	0.97	0.250 (0.018)	0.022	0.000	0.000	0.98	0.251 (0.021)	0.022	0.001	0.000	0.96
$D_{5,3} = 0.125$	0.124 (0.015)	0.016	-0.001	0.000	0.97	0.125 (0.014)	0.016	0.000	0.000	0.96	0.124 (0.014)	0.016	-0.001	0.000	0.99
$D_{4,4} = 0.090$	0.090 (0.007)	0.007	0.000	0.000	0.93	0.089 (0.007)	0.007	-0.001	0.000	0.99	0.090 (0.008)	0.008	0.000	0.000	0.95
$D_{5,5} = 0.250$	0.250 (0.020)	0.021	0.000	0.000	0.96	0.247 (0.019)	0.021	-0.003	0.000	0.94	0.247 (0.019)	0.022	-0.003	0.000	0.98
$D_{6,6} = 0.090$	0.090 (0.006)	0.007	0.000	0.000	0.99	0.090 (0.007)	0.007	0.000	0.000	0.95	0.089 (0.007)	0.008	-0.001	0.000	0.96
$\beta_{10} = 2.000$	1.999 (0.034)	0.037	-0.001	0.001	0.95	1.999 (0.039)	0.037	-0.001	0.002	0.94	2.002 (0.038)	0.037	0.002	0.001	0.94
$\beta_{11} = -0.100$	-0.101 (0.014)	0.015	-0.001	0.000	0.96	-0.104 (0.015)	0.016	-0.004	0.000	0.93	-0.110 (0.017)	0.017	-0.010	0.000	0.94
$\beta_{12} = 0.100$	0.100 (0.024)	0.026	0.000	0.001	0.97	0.095 (0.025)	0.026	-0.005	0.001	0.96	0.101 (0.025)	0.026	0.001	0.001	0.95
$\beta_{13} = -0.200$	-0.202 (0.054)	0.052	-0.002	0.003	0.94	-0.196 (0.050)	0.052	0.004	0.002	0.95	-0.203 (0.053)	0.053	-0.003	0.003	0.95
$\beta_{20} = -2.000$	-1.998 (0.033)	0.037	0.002	0.001	0.98	-2.001 (0.040)	0.037	-0.001	0.002	0.95	-1.998 (0.036)	0.038	0.002	0.001	0.98
$\beta_{21} = 0.100$	0.103 (0.014)	0.015	0.003	0.000	0.97	0.106 (0.015)	0.016	0.006	0.000	0.94	0.109 (0.015)	0.017	0.009	0.000	0.92
$\beta_{22} = -0.100$	-0.101 (0.026)	0.026	-0.001	0.001	0.96	-0.105 (0.024)	0.027	-0.005	0.001	0.99	-0.097 (0.027)	0.027	0.003	0.001	0.94
$\beta_{23} = 0.200$	0.192 (0.051)	0.052	-0.008	0.003	0.95	0.204 (0.055)	0.052	0.004	0.003	0.96	0.194 (0.056)	0.053	-0.006	0.003	0.93
$\beta_{30} = 2.000$	2.002 (0.036)	0.037	0.002	0.001	0.98	1.997 (0.039)	0.037	-0.003	0.001	0.95	2.000 (0.037)	0.037	0.000	0.001	0.95
$\beta_{31} = -0.100$	-0.101 (0.014)	0.015	-0.001	0.000	0.95	-0.103 (0.014)	0.016	-0.003	0.000	0.98	-0.108 (0.014)	0.017	-0.008	0.000	0.99
$\beta_{32} = 0.100$	0.098 (0.026)	0.026	-0.002	0.001	0.97	0.099 (0.021)	0.026	-0.001	0.000	0.98	0.102 (0.027)	0.026	0.002	0.001	0.93
$\beta_{33} = -0.200$	-0.205 (0.051)	0.052	-0.005	0.003	0.95	-0.195 (0.050)	0.052	0.005	0.003	0.95	-0.210 (0.052)	0.053	-0.010	0.003	0.96
$\sigma_1^2 = 0.160$	0.160 (0.004)	0.004	0.000	0.000	0.94	0.160 (0.004)	0.004	0.000	0.000	0.97	0.159 (0.004)	0.005	-0.001	0.000	0.97
$\sigma_2^2 = 0.160$	0.160 (0.004)	0.004	0.000	0.000	0.97	0.160 (0.005)	0.004	0.000	0.000	0.93	0.160 (0.004)	0.005	0.000	0.000	0.96
$\sigma_3^2 = 0.160$	0.160 (0.004)	0.004	0.000	0.000	0.95	0.160 (0.004)	0.004	0.000	0.000	0.95	0.160 (0.005)	0.005	0.000	0.000	0.95
$\gamma_1 = 0.500$	0.495 (0.132)	0.139	-0.005	0.017	0.96	0.502 (0.085)	0.090	0.002	0.007	0.94	0.510 (0.079)	0.080	0.010	0.006	0.97
$\gamma_2 = -0.500$	-0.503 (0.153)	0.138	-0.003	0.023	0.91	-0.502 (0.086)	0.090	-0.002	0.007	0.96	-0.509 (0.079)	0.080	-0.009	0.006	0.96
$\gamma_3 = 0.500$	0.499 (0.140)	0.138	-0.001	0.019	0.94	0.505 (0.089)	0.089	0.005	0.008	0.94	0.497 (0.076)	0.080	-0.003	0.006	0.95
$\zeta = -0.200$	-0.187 (0.290)	0.304	0.013	0.083	0.96	-0.198 (0.182)	0.180	0.002	0.033	0.95	-0.225 (0.143)	0.140	-0.025	0.021	0.92

Table B.8: Parameter estimates for differing failure rates $\omega \in \{10\%, 30\%, 50\%\}$. 'Mean (SD)' denotes the average estimated value with standard deviation of the parameter estimate. 'SE' denotes the mean standard error calculated at each model fit. Coverage probabilities are calculated from $\hat{\boldsymbol{\Omega}} \pm 1.96\text{SE}(\hat{\boldsymbol{\Omega}})$ Median [IQR] elapsed time taken for approximate EM algorithm to converge and standard error calculation for $\omega = 10\%$ was 5.390 [4.998, 5.604] seconds and total computation time was 9.182 [8.808, 9.364] seconds; for $\omega = 30\%$ was 8.337 [7.980, 8.697] seconds and total computation time was 12.023 [11.717, 12.454] seconds; for $\omega = 50\%$ was 16.566 [15.320, 18.506] seconds and total computation time was 20.099 [18.891, 22.085] seconds.

B.1.5 Random effects vech(D)

Additional results from simulation study undertaken in Section 3.4.6.

	$D^{(1)}$					$D^{(2)}$					$D^{(3)}$					$D^{(4)}$				
Parameter	Mean (SD)	SE	Bias	MSE	CP	Mean (SD)	SE	Bias	MSE	CP	Mean (SD)	SE	Bias	MSE	CP	Mean (SD)	SE	Bias	MSE	CP
$D_{1,1}$	0.750 (0.056)	0.055	0.000	0.003	0.94	2.487 (0.165)	0.173	-0.013	0.027	0.93	0.748 (0.058)	0.055	-0.002	0.003	0.94	2.484 (0.159)	0.173	-0.016	0.025	0.94
$D_{3,1}$	0.127 (0.042)	0.039	0.002	0.002	0.94	0.110 (0.109)	0.122	-0.015	0.012	0.97	0.135 (0.049)	0.050	0.010	0.003	0.96	0.135 (0.116)	0.138	0.010	0.013	0.98
$D_{5,1}$	0.124 (0.042)	0.039	-0.001	0.002	0.94	0.120 (0.121)	0.120	-0.005	0.015	0.94	0.127 (0.042)	0.044	0.002	0.002	0.96	0.105 (0.147)	0.148	-0.020	0.022	0.94
$D_{2,2}$	0.271 (0.018)	0.020	0.001	0.000	0.95	0.893 (0.061)	0.065	-0.007	0.004	0.97	0.134 (0.010)	0.010	-0.001	0.000	0.95	0.178 (0.014)	0.014	-0.002	0.000	0.94
$D_{3,3}$	0.747 (0.050)	0.055	-0.003	0.003	0.96	2.492 (0.144)	0.171	-0.008	0.021	0.99	1.246 (0.075)	0.088	-0.004	0.006	0.97	3.193 (0.226)	0.221	-0.057	0.054	0.95
$D_{5,3}$	0.127 (0.038)	0.039	0.002	0.001	0.93	0.104 (0.112)	0.122	-0.021	0.013	0.97	0.127 (0.055)	0.056	0.002	0.003	0.96	0.125 (0.157)	0.168	0.000	0.025	0.97
$D_{4,4}$	0.268 (0.019)	0.020	-0.002	0.000	0.95	0.889 (0.055)	0.064	-0.011	0.003	0.98	0.081 (0.006)	0.006	0.000	0.000	0.96	0.273 (0.019)	0.021	0.003	0.000	0.97
$D_{5,5}$	0.752 (0.051)	0.055	0.002	0.003	0.96	2.461 (0.168)	0.171	-0.039	0.029	0.94	0.991 (0.061)	0.071	-0.009	0.004	0.97	3.703 (0.259)	0.254	-0.047	0.068	0.94
$D_{6,6}$	0.269 (0.017)	0.020	-0.001	0.000	0.98	0.898 (0.054)	0.065	-0.002	0.003	0.98	0.108 (0.007)	0.008	0.000	0.000	0.99	0.365 (0.026)	0.027	0.005	0.001	0.99
$\beta_{10} = 2.000$	1.999 (0.057)	0.060	-0.001	0.003	0.97	1.989 (0.115)	0.105	-0.011	0.013	0.95	1.989 (0.056)	0.059	-0.011	0.003	0.95	2.001 (0.100)	0.105	0.001	0.010	0.96
$\beta_{11} = -0.100$	-0.105 (0.026)	0.025	-0.005	0.001	0.95	-0.117 (0.042)	0.043	-0.017	0.002	0.93	-0.103 (0.016)	0.018	-0.003	0.000	0.95	-0.108 (0.023)	0.021	-0.008	0.001	0.91
$\beta_{12} = 0.100$	0.102 (0.041)	0.042	0.002	0.002	0.95	0.099 (0.070)	0.071	-0.001	0.005	0.97	0.100 (0.042)	0.042	0.000	0.002	0.95	0.104 (0.081)	0.072	0.004	0.006	0.90
$\beta_{13} = -0.200$	-0.197 (0.088)	0.085	0.003	0.008	0.92	-0.198 (0.153)	0.151	0.002	0.023	0.97	-0.194 (0.086)	0.085	0.006	0.007	0.93	-0.198 (0.132)	0.150	0.002	0.017	0.97
$\beta_{20} = -2.000$	-2.001 (0.069)	0.060	-0.001	0.005	0.90	-2.001 (0.095)	0.105	-0.001	0.009	0.97	-2.005 (0.076)	0.075	-0.005	0.006	0.94	-2.005 (0.128)	0.119	-0.005	0.016	0.93
$\beta_{21} = 0.100$	0.108 (0.024)	0.025	0.008	0.001	0.94	0.104 (0.044)	0.043	0.004	0.002	0.92	0.103 (0.014)	0.015	0.003	0.000	0.98	0.104 (0.027)	0.026	0.004	0.001	0.93
$\beta_{22} = -0.100$	-0.099 (0.045)	0.042	0.001	0.002	0.94	-0.093 (0.074)	0.073	0.007	0.006	0.96	-0.104 (0.040)	0.053	-0.004	0.002	0.99	-0.110 (0.075)	0.080	-0.010	0.006	0.94
$\beta_{23} = 0.200$	0.196 (0.088)	0.085	-0.004	0.008	0.94	0.197 (0.144)	0.151	-0.003	0.021	0.96	0.200 (0.111)	0.107	0.000	0.012	0.95	0.204 (0.183)	0.170	0.004	0.033	0.94
$\beta_{30} = 2.000$	1.999 (0.060)	0.060	-0.001	0.004	0.94	1.999 (0.100)	0.104	-0.001	0.010	0.94	1.991 (0.067)	0.068	-0.009	0.005	0.94	2.015 (0.123)	0.128	0.015	0.015	0.98
$\beta_{31} = -0.100$	-0.107 (0.024)	0.025	-0.007	0.001	0.93	-0.113 (0.044)	0.043	-0.013	0.002	0.92	-0.100 (0.014)	0.017	0.000	0.000	0.96	-0.109 (0.025)	0.029	-0.009	0.001	0.96
$\beta_{32} = 0.100$	0.102 (0.042)	0.042	0.002	0.002	0.94	0.102 (0.065)	0.071	0.002	0.004	0.98	0.098 (0.047)	0.048	-0.002	0.002	0.97	0.098 (0.086)	0.084	-0.002	0.007	0.92
$\beta_{33} = -0.200$	-0.199 (0.079)	0.085	0.001	0.006	0.98	-0.200 (0.132)	0.150	0.000	0.017	0.96	-0.193 (0.104)	0.097	0.007	0.011	0.91	-0.215 (0.166)	0.182	-0.015	0.028	0.98
$\sigma_1^2 = 0.160$	0.160 (0.004)	0.004	0.000	0.000	0.98	0.159 (0.004)	0.004	-0.001	0.000	0.94	0.160 (0.004)	0.004	0.000	0.000	0.97	0.160 (0.004)	0.004	0.000	0.000	0.96
$\sigma_2^2 = 0.160$	0.159 (0.003)	0.004	-0.001	0.000	0.97	0.159 (0.005)	0.004	-0.001	0.000	0.92	0.159 (0.004)	0.004	-0.001	0.000	0.98	0.160 (0.004)	0.004	0.000	0.000	0.97
$\sigma_3^2 = 0.160$	0.160 (0.004)	0.004	0.000	0.000	0.95	0.160 (0.004)	0.004	0.000	0.000	0.98	0.160 (0.004)	0.004	0.000	0.000	0.97	0.160 (0.004)	0.004	0.000	0.000	0.99
$\gamma_1 = 0.500$	0.506 (0.070)	0.053	0.006	0.005	0.87	0.511 (0.034)	0.032	0.011	0.001	0.93	0.491 (0.071)	0.072	-0.009	0.005	0.96	0.504 (0.058)	0.049	0.004	0.003	0.89
$\gamma_2 = -0.500$	-0.518 (0.061)	0.054	-0.018	0.004	0.91	-0.510 (0.033)	0.032	-0.010	0.001	0.93	-0.501 (0.074)	0.074	-0.001	0.005	0.93	-0.508 (0.041)	0.041	-0.008	0.002	0.94
$\gamma_3 = 0.500$	0.509 (0.060)	0.054	0.009	0.004	0.91	0.513 (0.036)	0.032	0.013	0.001	0.92	0.506 (0.076)	0.073	0.006	0.006	0.95	0.508 (0.043)	0.037	0.008	0.002	0.91
$\zeta = -0.200$	-0.182 (0.183)	0.213	0.018	0.034	0.96	-0.178 (0.228)	0.217	0.022	0.052	0.93	-0.195 (0.244)	0.227	0.005	0.059	0.92	-0.232 (0.207)	0.233	-0.032	0.044	0.99

Table B.9: Parameter estimates for different variance covariance matrices D. 'Mean (SD)' denotes the average estimated value with standard deviation of the parameter estimate. 'SE' denotes the mean standard error calculated at each model fit. Coverage probabilities are calculated from $\hat{\boldsymbol{\Omega}} \pm 1.96\text{SE}(\hat{\boldsymbol{\Omega}})$. Median [IQR] elapsed time taken for approximate EM algorithm to converge and standard error calculation for $D^{(1)}$ was 4.888 [4.599, 5.334] seconds and total computation time was 8.078 [7.788, 8.524] seconds; for $D^{(2)}$ was 6.041 [5.718, 7.133] seconds and total computation time was 9.357 [9.011, 10.452] seconds; for $D^{(3)}$ was 4.342 [4.181, 4.579] seconds and total computation time was 7.704 [7.547, 7.902] seconds; for $D^{(4)}$ was 5.156 [4.920, 5.409] seconds and total computation time was 8.535 [8.288, 8.755] seconds. Target values for each of the four D matrices are given in Section 3.4.6.

B.1.6 Censoring rate Υ

Additional results from simulation study undertaken in Section 3.6.2. Lowest censoring occurs ($\approx 13\%$) for $\Upsilon = e^{-3.5}$ and highest ($\approx 50\%$) at $\Upsilon = e^{-1.9}$.

	Lowest censoring					Medium censoring					Highest censoring				
Parameter	Mean (SD)	SE	Bias	MSE	CP	Mean (SD)	SE	Bias	MSE	CP	Mean (SD)	SE	Bias	MSE	CP
$D_{1,1} = 0.250$	0.247 (0.019)	0.021	-0.003	0.000	0.94	0.249 (0.020)	0.022	-0.001	0.000	0.95	0.248 (0.018)	0.022	-0.002	0.000	0.98
$D_{3,1} = 0.125$	0.123 (0.015)	0.016	-0.002	0.000	0.96	0.125 (0.015)	0.016	0.000	0.000	1.00	0.123 (0.013)	0.017	-0.002	0.000	0.99
$D_{5,1} = 0.125$	0.122 (0.016)	0.016	-0.003	0.000	0.93	0.124 (0.016)	0.016	-0.001	0.000	0.95	0.126 (0.016)	0.017	0.001	0.000	0.96
$D_{2,2} = 0.090$	0.090 (0.006)	0.007	0.000	0.000	0.98	0.090 (0.006)	0.008	0.000	0.000	0.97	0.089 (0.008)	0.008	-0.001	0.000	0.96
$D_{3,3} = 0.250$	0.250 (0.017)	0.021	0.000	0.000	0.97	0.252 (0.022)	0.022	0.002	0.000	0.96	0.253 (0.020)	0.023	0.003	0.000	0.97
$D_{5,3} = 0.125$	0.125 (0.016)	0.016	0.000	0.000	0.93	0.124 (0.013)	0.016	-0.001	0.000	0.97	0.125 (0.015)	0.017	0.000	0.000	0.97
$D_{4,4} = 0.090$	0.090 (0.007)	0.007	0.000	0.000	0.94	0.089 (0.006)	0.007	-0.001	0.000	0.96	0.089 (0.008)	0.008	-0.001	0.000	0.97
$D_{5,5} = 0.250$	0.247 (0.022)	0.021	-0.003	0.000	0.90	0.249 (0.018)	0.021	-0.001	0.000	0.99	0.251 (0.022)	0.022	0.001	0.000	0.95
$D_{6,6} = 0.090$	0.089 (0.006)	0.007	-0.001	0.000	0.99	0.089 (0.007)	0.007	-0.001	0.000	0.97	0.089 (0.008)	0.008	-0.001	0.000	0.94
$\beta_{10} = 2.000$	1.998 (0.039)	0.037	-0.002	0.001	0.94	1.997 (0.037)	0.037	-0.003	0.001	0.95	1.995 (0.034)	0.038	-0.005	0.001	0.97
$\beta_{11} = -0.100$	-0.104 (0.013)	0.015	-0.004	0.000	0.99	-0.101 (0.017)	0.016	-0.001	0.000	0.92	-0.103 (0.017)	0.017	-0.003	0.000	0.96
$\beta_{12} = 0.100$	0.100 (0.023)	0.026	0.000	0.001	0.99	0.100 (0.026)	0.027	0.000	0.001	0.96	0.108 (0.026)	0.027	0.008	0.001	0.93
$\beta_{13} = -0.200$	-0.199 (0.053)	0.052	0.001	0.003	0.95	-0.199 (0.060)	0.053	0.001	0.004	0.92	-0.195 (0.046)	0.053	0.005	0.002	0.98
$\beta_{20} = -2.000$	-2.004 (0.040)	0.037	-0.004	0.002	0.95	-1.998 (0.039)	0.037	0.002	0.002	0.94	-2.002 (0.036)	0.038	-0.002	0.001	0.96
$\beta_{21} = 0.100$	0.102 (0.013)	0.015	0.002	0.000	0.99	0.101 (0.018)	0.016	0.001	0.000	0.91	0.104 (0.016)	0.017	0.004	0.000	0.95
$\beta_{22} = -0.100$	-0.102 (0.024)	0.026	-0.002	0.001	0.97	-0.096 (0.026)	0.027	0.004	0.001	0.93	-0.098 (0.024)	0.027	0.002	0.001	0.98
$\beta_{23} = 0.200$	0.205 (0.053)	0.052	0.005	0.003	0.94	0.204 (0.057)	0.053	0.004	0.003	0.93	0.203 (0.050)	0.053	0.003	0.002	0.97
$\beta_{30} = 2.000$	1.999 (0.034)	0.037	-0.001	0.001	0.97	1.996 (0.033)	0.038	-0.004	0.001	0.97	2.005 (0.034)	0.038	0.005	0.001	0.97
$\beta_{31} = -0.100$	-0.101 (0.013)	0.015	-0.001	0.000	0.96	-0.101 (0.015)	0.016	-0.001	0.000	0.96	-0.105 (0.015)	0.017	-0.005	0.000	0.97
$\beta_{32} = 0.100$	0.099 (0.022)	0.026	-0.001	0.000	0.96	0.098 (0.025)	0.027	-0.002	0.001	0.93	0.106 (0.023)	0.027	0.006	0.001	0.98
$\beta_{33} = -0.200$	-0.203 (0.053)	0.052	-0.003	0.003	0.94	-0.196 (0.047)	0.053	0.004	0.002	0.98	-0.197 (0.045)	0.053	0.003	0.002	0.98
$\sigma_1^2 = 0.160$	0.160 (0.004)	0.004	0.000	0.000	0.92	0.160 (0.004)	0.004	0.000	0.000	0.98	0.160 (0.004)	0.005	0.000	0.000	0.98
$\sigma_2^2 = 0.160$	0.160 (0.004)	0.004	0.000	0.000	0.96	0.160 (0.004)	0.004	0.000	0.000	0.97	0.159 (0.005)	0.005	-0.001	0.000	0.95
$\sigma_3^2 = 0.160$	0.160 (0.004)	0.004	0.000	0.000	0.98	0.160 (0.004)	0.004	0.000	0.000	0.98	0.159 (0.005)	0.005	-0.001	0.000	0.96
$\gamma_1 = 0.500$	0.508 (0.100)	0.105	0.008	0.010	0.92	0.490 (0.128)	0.116	-0.010	0.016	0.93	0.517 (0.148)	0.139	0.017	0.022	0.96
$\gamma_2 = -0.500$	-0.534 (0.097)	0.104	-0.034	0.010	0.94	-0.498 (0.114)	0.115	0.002	0.013	0.93	-0.533 (0.160)	0.137	-0.033	0.027	0.91
$\gamma_3 = 0.500$	0.507 (0.118)	0.106	0.007	0.014	0.92	0.497 (0.117)	0.114	-0.003	0.014	0.99	0.536 (0.152)	0.136	0.036	0.024	0.91
$\zeta = -0.200$	-0.211 (0.236)	0.236	-0.011	0.055	0.95	-0.208 (0.244)	0.251	-0.008	0.059	0.97	-0.237 (0.292)	0.283	-0.037	0.086	0.95

Table B.10: Parameter estimates for differing censoring rates $\Upsilon \in \{e^{-3.5}, e^{-2.6}, e^{-1.9}\}$ as outlined in Section 3.6.2. 'Mean (SD)' denotes the average estimated value with standard deviation of the parameter estimate. 'SE' denotes the mean standard error calculated at each model fit. Coverage probabilities are calculated from $\hat{\boldsymbol{\Omega}} \pm 1.96\text{SE}(\hat{\boldsymbol{\Omega}})$. Median [IQR] elapsed time taken for approximate EM algorithm to converge and standard error calculation for lowest censoring was 7.037 [6.756, 7.319] seconds and total computation time was 10.165 [9.796, 10.487] seconds; for medium censoring was 6.747 [6.430, 7.121] seconds and total computation time was 9.663 [9.383, 10.148] seconds; for highest censoring was 6.765 [6.358, 7.181] seconds and total computation time was 9.593 [9.136, 9.974] seconds.

B.2 Additional results from Chapter 4

Additional results from the simulations carried out in Chapter 4 are presented here. Provided summary measures are identical to (B.1).

B.2.1 Mixed trivariate simulation

Additional results from simulation study undertaken in Section 4.5.3.

Parameter	Mean (SD)	SE	Bias	MSE	CP
$D_{1,1} = 0.250$	0.246 (0.021)	0.021	-0.004	0.000	0.93
$D_{3,1} = 0.125$	0.124 (0.020)	0.020	-0.001	0.000	0.94
$D_{5,1} = 0.125$	0.117 (0.044)	0.047	-0.008	0.002	0.96
$D_{2,2} = 0.090$	0.090 (0.007)	0.007	0.000	0.000	0.96
$D_{3,3} = 0.500$	0.487 (0.036)	0.038	-0.013	0.001	0.93
$D_{5,3} = 0.125$	0.124 (0.057)	0.062	-0.001	0.003	0.96
$D_{4,4} = 0.090$	0.088 (0.007)	0.008	-0.002	0.000	0.96
$D_{5,5} = 2.000$	1.712 (0.211)	0.260	-0.288	0.127	0.78
$\beta_{10} = -2.000$	-2.013 (0.037)	0.037	-0.013	0.002	0.93
$\beta_{11} = 0.100$	0.099 (0.016)	0.016	-0.001	0.000	0.95
$\beta_{12} = -0.100$	-0.102 (0.027)	0.026	-0.002	0.001	0.95
$\beta_{13} = 0.200$	0.204 (0.048)	0.052	0.004	0.002	0.95
$\beta_{20} = 2.000$	2.006 (0.049)	0.049	0.006	0.002	0.95
$\beta_{21} = -0.100$	-0.095 (0.015)	0.016	0.005	0.000	0.96
$\beta_{22} = 0.100$	0.097 (0.035)	0.035	-0.003	0.001	0.95
$\beta_{23} = 0.200$	0.200 (0.070)	0.069	0.000	0.005	0.96
$\beta_{30} = 1.000$	0.895 (0.132)	0.124	-0.105	0.028	0.85
$\beta_{31} = -1.000$	-0.993 (0.044)	0.047	0.007	0.002	0.96
$\beta_{32} = 1.000$	1.044 (0.095)	0.086	0.044	0.011	0.90
$\beta_{33} = -1.000$	-1.059 (0.166)	0.156	-0.059	0.031	0.92
$\sigma_1^2 = 0.160$	0.160 (0.003)	0.004	0.000	0.000	0.99
$\gamma_1 = 0.500$	0.499 (0.099)	0.094	-0.001	0.010	0.96
$\gamma_2 = -0.500$	-0.490 (0.087)	0.087	0.010	0.008	0.95
$\gamma_3 = 0.500$	0.464 (0.089)	0.094	-0.036	0.009	0.94
$\zeta = -0.200$	-0.223 (0.187)	0.182	-0.023	0.035	0.93

Table B.11: Parameter estimates for the trivariate joint model simulations. Median [IQR] elapsed time taken for approximate EM algorithm to converge and standard error calculation for was 10.359 [9.593, 11.160] seconds and total computation time was 14.893 [14.079, 15.594] seconds

B.2.2 Further trivariate simulations

Additional results from the simulation study undertaken in Section 4.5.4

	$r = 5$					$r = 10$					$r = 15$				
Parameter	Emp. Mean (SD)	Mean SE	Bias	MSE	CP	Emp. Mean (SD)	Mean SE	Bias	MSE	CP	Emp. Mean (SD)	Mean SE	Bias	MSE	CP
$D_{1,1} = 0.250$	0.248 (0.031)	0.035	-0.002	0.001	0.95	0.249 (0.027)	0.031	-0.001	0.001	0.97	0.247 (0.028)	0.029	-0.003	0.001	0.94
$D_{3,1} = 0.250$	0.244 (0.031)	0.036	-0.006	0.001	0.97	0.247 (0.030)	0.034	-0.003	0.001	0.97	0.246 (0.030)	0.033	-0.004	0.001	0.95
$D_{5,1} = 0.250$	0.233 (0.081)	0.086	-0.017	0.007	0.95	0.240 (0.057)	0.072	-0.010	0.003	0.98	0.234 (0.061)	0.065	-0.016	0.004	0.94
$D_{2,2} = 0.090$	0.090 (0.009)	0.011	0.000	0.000	0.98	0.090 (0.010)	0.010	0.000	0.000	0.95	0.090 (0.009)	0.010	0.000	0.000	0.98
$D_{3,3} = 0.500$	0.482 (0.055)	0.060	-0.018	0.003	0.92	0.487 (0.055)	0.057	-0.013	0.003	0.92	0.487 (0.051)	0.055	-0.013	0.003	0.95
$D_{5,3} = 0.250$	0.233 (0.100)	0.108	-0.017	0.010	0.97	0.242 (0.080)	0.094	-0.008	0.006	0.97	0.234 (0.081)	0.087	-0.016	0.007	0.95
$D_{4,4} = 0.100$	0.093 (0.011)	0.012	-0.007	0.000	0.88	0.095 (0.010)	0.011	-0.005	0.000	0.92	0.098 (0.010)	0.011	-0.002	0.000	0.95
$D_{5,5} = 2.000$	1.664 (0.435)	0.540	-0.336	0.301	0.88	1.803 (0.294)	0.373	-0.197	0.125	0.94	1.802 (0.265)	0.316	-0.198	0.109	0.90
$\beta_{10} = 2.000$	1.980 (0.057)	0.057	-0.020	0.004	0.94	1.990 (0.052)	0.054	-0.010	0.003	0.93	1.996 (0.049)	0.052	-0.004	0.002	0.96
$\beta_{11} = -0.100$	-0.103 (0.021)	0.023	-0.003	0.000	0.97	-0.098 (0.022)	0.022	0.002	0.000	0.95	-0.099 (0.021)	0.022	0.001	0.000	0.95
$\beta_{12} = 0.100$	0.097 (0.035)	0.041	-0.003	0.001	0.96	0.103 (0.033)	0.039	0.003	0.001	0.97	0.102 (0.031)	0.037	0.002	0.001	0.98
$\beta_{13} = -0.200$	-0.198 (0.075)	0.080	0.002	0.006	0.95	-0.206 (0.070)	0.076	-0.006	0.005	0.97	-0.209 (0.074)	0.073	-0.009	0.006	0.94
$\beta_{20} = 2.000$	2.004 (0.076)	0.074	0.004	0.006	0.94	2.001 (0.067)	0.072	0.001	0.005	0.96	2.009 (0.062)	0.071	0.009	0.004	0.96
$\beta_{21} = -0.100$	-0.097 (0.026)	0.025	0.003	0.001	0.93	-0.096 (0.022)	0.024	0.004	0.000	0.97	-0.095 (0.022)	0.023	0.005	0.001	0.96
$\beta_{22} = 0.100$	0.096 (0.049)	0.053	-0.004	0.002	0.96	0.102 (0.047)	0.052	0.002	0.002	0.97	0.100 (0.045)	0.051	0.000	0.002	0.96
$\beta_{23} = -0.200$	-0.196 (0.100)	0.104	0.004	0.010	0.96	-0.201 (0.099)	0.101	-0.001	0.010	0.95	-0.203 (0.096)	0.100	-0.003	0.009	0.96
$\beta_{30} = 1.000$	0.924 (0.217)	0.217	-0.076	0.053	0.93	0.877 (0.203)	0.180	-0.123	0.056	0.89	0.861 (0.185)	0.162	-0.139	0.054	0.82
$\beta_{31} = -1.000$	-0.986 (0.074)	0.089	0.014	0.006	0.97	-0.985 (0.057)	0.062	0.015	0.004	0.94	-0.988 (0.050)	0.051	0.012	0.003	0.94
$\beta_{32} = 1.000$	1.005 (0.161)	0.154	0.005	0.026	0.94	1.056 (0.136)	0.126	0.056	0.021	0.93	1.057 (0.120)	0.114	0.057	0.018	0.91
$\beta_{33} = -1.000$	-1.024 (0.278)	0.270	-0.024	0.077	0.94	-1.060 (0.256)	0.228	-0.060	0.069	0.91	-1.050 (0.234)	0.209	-0.050	0.057	0.92
$\sigma_1^2 = 0.160$	0.159 (0.009)	0.010	-0.001	0.000	0.96	0.160 (0.005)	0.006	0.000	0.000	0.98	0.160 (0.004)	0.005	0.000	0.000	0.97
$\gamma_1 = 0.500$	0.520 (0.242)	0.244	0.020	0.059	0.92	0.516 (0.211)	0.231	0.016	0.045	0.95	0.512 (0.239)	0.224	0.012	0.057	0.92
$\gamma_2 = -0.500$	-0.540 (0.206)	0.220	-0.040	0.044	0.95	-0.493 (0.198)	0.203	0.007	0.039	0.95	-0.502 (0.190)	0.195	-0.002	0.036	0.96
$\gamma_3 = 0.500$	0.495 (0.238)	0.267	-0.005	0.056	0.97	0.494 (0.227)	0.208	-0.006	0.051	0.94	0.485 (0.208)	0.194	-0.015	0.043	0.93
$\zeta = -0.200$	-0.197 (0.442)	0.452	0.003	0.194	0.93	-0.240 (0.427)	0.454	-0.040	0.183	0.97	-0.180 (0.454)	0.452	0.020	0.205	0.96

Table B.12: Parameter estimates for $\omega = 10\%$ for differing maximal longitudinal profile lengths r. 'Emp. Mean (SD)' denotes the average estimated value with the standard deviation of parameter estimates and Mean SE the mean standard error calculated for each parameter from each model fit. Coverage probabilities are calculated from $\hat{\boldsymbol{\Omega}} \pm 1.96\text{SE}(\hat{\boldsymbol{\Omega}})$. The median [IQR] elapsed times for the approximate EM algorithm to converge and standard errors calculated was 4.628 [4.125, 5.237] seconds for $r = 5$, 3.217 [2.896, 3.608] seconds for $r = 10$ and 3.175 [2.828, 3.631] seconds for $r = 15$. *Total* computation time was 6.585 [6.112, 7.258] seconds for $r = 5$, 5.844 [5.522, 6.336] seconds for $r = 10$ and 6.511 [6.175, 6.970] seconds for $r = 15$.

	$r = 5$					$r = 10$					$r = 15$				
Parameter	Emp. Mean (SD)	Mean SE	Bias	MSE	CP	Emp. Mean (SD)	Mean SE	Bias	MSE	CP	Emp. Mean (SD)	Mean SE	Bias	MSE	CP
$D_{1,1} = 0.250$	0.249 (0.031)	0.037	-0.001	0.001	0.96	0.254 (0.030)	0.032	0.004	0.001	0.96	0.248 (0.027)	0.030	-0.002	0.001	0.97
$D_{3,1} = 0.250$	0.241 (0.031)	0.037	-0.009	0.001	0.97	0.250 (0.032)	0.035	0.000	0.001	0.94	0.245 (0.032)	0.033	-0.005	0.001	0.94
$D_{5,1} = 0.250$	0.229 (0.074)	0.089	-0.021	0.006	0.98	0.231 (0.066)	0.073	-0.019	0.005	0.94	0.233 (0.058)	0.067	-0.017	0.004	0.95
$D_{2,2} = 0.090$	0.089 (0.010)	0.012	-0.001	0.000	0.96	0.091 (0.010)	0.011	0.001	0.000	0.95	0.090 (0.009)	0.011	0.000	0.000	0.95
$D_{3,3} = 0.500$	0.471 (0.053)	0.060	-0.029	0.004	0.93	0.490 (0.052)	0.058	-0.010	0.003	0.94	0.484 (0.048)	0.054	-0.016	0.003	0.94
$D_{5,3} = 0.250$	0.233 (0.103)	0.111	-0.017	0.011	0.96	0.247 (0.089)	0.095	-0.003	0.008	0.96	0.236 (0.076)	0.088	-0.014	0.006	0.97
$D_{4,4} = 0.100$	0.095 (0.012)	0.013	-0.005	0.000	0.92	0.096 (0.012)	0.012	-0.004	0.000	0.95	0.098 (0.011)	0.012	-0.002	0.000	0.94
$D_{5,5} = 2.000$	1.687 (0.448)	0.603	-0.313	0.297	0.92	1.696 (0.287)	0.384	-0.304	0.175	0.89	1.766 (0.266)	0.336	-0.234	0.125	0.90
$\beta_{10} = 2.000$	1.984 (0.054)	0.057	-0.016	0.003	0.96	1.982 (0.049)	0.055	-0.018	0.003	0.96	1.991 (0.046)	0.052	-0.009	0.002	0.97
$\beta_{11} = -0.100$	-0.105 (0.025)	0.025	-0.005	0.001	0.94	-0.103 (0.021)	0.024	-0.003	0.000	0.98	-0.102 (0.022)	0.023	-0.002	0.000	0.95
$\beta_{12} = 0.100$	0.108 (0.037)	0.041	0.008	0.001	0.96	0.104 (0.033)	0.039	0.004	0.001	0.98	0.104 (0.038)	0.038	0.004	0.001	0.97
$\beta_{13} = -0.200$	-0.201 (0.078)	0.081	-0.001	0.006	0.95	-0.197 (0.070)	0.077	0.003	0.005	0.98	-0.209 (0.076)	0.074	-0.009	0.006	0.94
$\beta_{20} = 2.000$	2.016 (0.071)	0.073	0.016	0.005	0.95	2.000 (0.068)	0.072	0.000	0.005	0.96	2.006 (0.068)	0.071	0.006	0.005	0.95
$\beta_{21} = -0.100$	-0.094 (0.026)	0.027	0.006	0.001	0.94	-0.096 (0.023)	0.025	0.004	0.001	0.94	-0.095 (0.023)	0.024	0.005	0.001	0.96
$\beta_{22} = 0.100$	0.106 (0.045)	0.052	0.006	0.002	0.98	0.103 (0.045)	0.052	0.003	0.002	0.97	0.106 (0.050)	0.051	0.006	0.002	0.94
$\beta_{23} = -0.200$	-0.195 (0.107)	0.103	0.005	0.011	0.97	-0.199 (0.093)	0.102	0.001	0.009	0.96	-0.210 (0.096)	0.100	-0.010	0.009	0.96
$\beta_{30} = 1.000$	0.913 (0.232)	0.222	-0.087	0.061	0.90	0.894 (0.199)	0.182	-0.106	0.051	0.86	0.883 (0.181)	0.165	-0.117	0.046	0.83
$\beta_{31} = -1.000$	-1.000 (0.092)	0.102	0.000	0.008	0.98	-0.997 (0.063)	0.070	0.003	0.004	0.97	-0.997 (0.050)	0.058	0.003	0.002	0.97
$\beta_{32} = 1.000$	0.995 (0.153)	0.161	-0.005	0.023	0.96	1.022 (0.123)	0.127	0.022	0.016	0.95	1.065 (0.128)	0.116	0.065	0.021	0.89
$\beta_{33} = -1.000$	-1.006 (0.285)	0.281	-0.006	0.081	0.95	-1.033 (0.240)	0.230	-0.033	0.058	0.92	-1.040 (0.240)	0.212	-0.040	0.059	0.91
$\sigma_1^2 = 0.160$	0.160 (0.010)	0.011	0.000	0.000	0.96	0.160 (0.006)	0.006	0.000	0.000	0.95	0.160 (0.005)	0.005	0.000	0.000	0.96
$\gamma_1 = 0.500$	0.513 (0.149)	0.160	0.013	0.022	0.96	0.507 (0.132)	0.146	0.007	0.017	0.96	0.507 (0.143)	0.143	0.007	0.020	0.94
$\gamma_2 = -0.500$	-0.495 (0.129)	0.147	0.005	0.017	0.98	-0.501 (0.124)	0.131	-0.001	0.015	0.95	-0.489 (0.127)	0.127	0.011	0.016	0.96
$\gamma_3 = 0.500$	0.462 (0.148)	0.185	-0.038	0.023	0.95	0.481 (0.126)	0.143	-0.019	0.016	0.96	0.481 (0.129)	0.127	-0.019	0.017	0.93
$\zeta = -0.200$	-0.205 (0.257)	0.262	-0.005	0.066	0.96	-0.201 (0.238)	0.260	-0.001	0.056	0.97	-0.226 (0.254)	0.260	-0.026	0.065	0.96

Table B.13: Parameter estimates for $\omega = 30\%$ for differing maximal longitudinal profile lengths r. 'Emp. Mean (SD)' denotes the average estimated value with the standard deviation of parameter estimates and Mean SE the mean standard error calculated for each parameter from each model fit. Coverage probabilities are calculated from $\hat{\boldsymbol{\Omega}} \pm 1.96\text{SE}(\hat{\boldsymbol{\Omega}})$. The median [IQR] elapsed times for the approximate EM algorithm to converge and standard errors calculated was 5.868 [5.167, 6.916] seconds for $r = 5$, 4.110 [3.720, 4.577] seconds for $r = 10$ and 3.964 [3.620, 4.383] seconds for $r = 15$. *Total* computation time was 7.777 [7.082, 8.809] seconds for $r = 5$, 6.716 [6.296, 7.201] seconds for $r = 10$ and 7.240 [6.822, 7.716] seconds for $r = 15$.

	$r = 5$					$r = 10$					$r = 15$				
Parameter	Emp. Mean (SD)	Mean SE	Bias	MSE	CP	Emp. Mean (SD)	Mean SE	Bias	MSE	CP	Emp. Mean (SD)	Mean SE	Bias	MSE	CP
$D_{1,1} = 0.250$	0.251 (0.031)	0.038	0.001	0.001	0.97	0.246 (0.028)	0.033	-0.004	0.001	0.98	0.247 (0.026)	0.030	-0.003	0.001	0.97
$D_{3,1} = 0.250$	0.241 (0.033)	0.037	-0.009	0.001	0.94	0.243 (0.032)	0.035	-0.007	0.001	0.95	0.245 (0.031)	0.034	-0.005	0.001	0.95
$D_{5,1} = 0.250$	0.224 (0.075)	0.090	-0.026	0.006	0.97	0.229 (0.062)	0.074	-0.021	0.004	0.96	0.233 (0.060)	0.068	-0.017	0.004	0.98
$D_{2,2} = 0.090$	0.091 (0.013)	0.014	0.001	0.000	0.94	0.089 (0.011)	0.012	-0.001	0.000	0.97	0.089 (0.011)	0.012	-0.001	0.000	0.97
$D_{3,3} = 0.500$	0.470 (0.055)	0.061	-0.030	0.004	0.91	0.482 (0.050)	0.057	-0.018	0.003	0.95	0.482 (0.049)	0.056	-0.018	0.003	0.94
$D_{5,3} = 0.250$	0.228 (0.101)	0.112	-0.022	0.011	0.95	0.238 (0.081)	0.095	-0.012	0.007	0.97	0.246 (0.078)	0.089	-0.004	0.006	0.99
$D_{4,4} = 0.100$	0.095 (0.014)	0.015	-0.005	0.000	0.93	0.097 (0.013)	0.014	-0.003	0.000	0.94	0.096 (0.012)	0.013	-0.004	0.000	0.93
$D_{5,5} = 2.000$	1.543 (0.402)	0.618	-0.457	0.369	0.92	1.608 (0.317)	0.403	-0.392	0.253	0.81	1.686 (0.259)	0.350	-0.314	0.166	0.83
$\beta_{10} = 2.000$	1.986 (0.056)	0.058	-0.014	0.003	0.95	1.982 (0.053)	0.054	-0.018	0.003	0.94	1.988 (0.054)	0.053	-0.012	0.003	0.92
$\beta_{11} = -0.100$	-0.105 (0.025)	0.027	-0.005	0.001	0.96	-0.104 (0.025)	0.025	-0.004	0.001	0.95	-0.107 (0.025)	0.024	-0.007	0.001	0.95
$\beta_{12} = 0.100$	0.104 (0.038)	0.042	0.004	0.001	0.95	0.099 (0.034)	0.039	-0.001	0.001	0.96	0.098 (0.034)	0.038	-0.002	0.001	0.98
$\beta_{13} = -0.200$	-0.204 (0.080)	0.082	-0.004	0.006	0.95	-0.195 (0.077)	0.077	0.005	0.006	0.94	-0.206 (0.079)	0.075	-0.006	0.006	0.93
$\beta_{20} = 2.000$	2.023 (0.072)	0.073	0.023	0.006	0.94	1.998 (0.069)	0.072	-0.002	0.005	0.95	2.007 (0.071)	0.071	0.007	0.005	0.95
$\beta_{21} = -0.100$	-0.085 (0.026)	0.029	0.015	0.001	0.92	-0.088 (0.026)	0.027	0.012	0.001	0.93	-0.091 (0.023)	0.026	0.009	0.001	0.97
$\beta_{22} = 0.100$	0.100 (0.050)	0.053	0.000	0.002	0.95	0.098 (0.050)	0.052	-0.002	0.002	0.96	0.096 (0.049)	0.051	-0.004	0.002	0.95
$\beta_{23} = -0.200$	-0.202 (0.101)	0.104	-0.002	0.010	0.95	-0.186 (0.103)	0.102	0.014	0.011	0.95	-0.202 (0.097)	0.100	-0.002	0.009	0.94
$\beta_{30} = 1.000$	0.909 (0.213)	0.226	-0.091	0.053	0.94	0.885 (0.188)	0.185	-0.115	0.048	0.91	0.872 (0.205)	0.168	-0.128	0.058	0.82
$\beta_{31} = -1.000$	-1.004 (0.088)	0.113	-0.004	0.008	0.98	-1.005 (0.069)	0.079	-0.005	0.005	0.98	-1.003 (0.062)	0.066	-0.003	0.004	0.96
$\beta_{32} = 1.000$	1.004 (0.156)	0.169	0.004	0.024	0.96	1.024 (0.137)	0.131	0.024	0.019	0.94	1.037 (0.119)	0.118	0.037	0.015	0.96
$\beta_{33} = -1.000$	-1.014 (0.297)	0.290	-0.014	0.088	0.94	-1.019 (0.248)	0.235	-0.019	0.062	0.93	-1.049 (0.243)	0.216	-0.049	0.061	0.91
$\sigma_1^2 = 0.160$	0.159 (0.010)	0.012	-0.001	0.000	0.97	0.160 (0.006)	0.007	0.000	0.000	0.98	0.160 (0.005)	0.005	0.000	0.000	0.96
$\gamma_1 = 0.500$	0.494 (0.133)	0.151	-0.006	0.018	0.97	0.510 (0.119)	0.139	0.010	0.014	0.97	0.514 (0.129)	0.133	0.014	0.017	0.95
$\gamma_2 = -0.500$	-0.466 (0.113)	0.141	0.034	0.014	0.97	-0.479 (0.098)	0.124	0.021	0.010	0.99	-0.488 (0.108)	0.120	0.012	0.012	0.96
$\gamma_3 = 0.500$	0.452 (0.127)	0.183	-0.048	0.018	0.95	0.475 (0.108)	0.138	-0.025	0.012	0.98	0.465 (0.110)	0.119	-0.035	0.013	0.96
$\zeta = -0.200$	-0.247 (0.225)	0.214	-0.047	0.053	0.94	-0.222 (0.207)	0.215	-0.022	0.043	0.96	-0.216 (0.212)	0.214	-0.016	0.045	0.95

Table B.14: Parameter estimates for $\omega = 50\%$ for differing maximal longitudinal profile lengths r. 'Emp. Mean (SD)' denotes the average estimated value with the standard deviation of parameter estimates and Mean SE the mean standard error calculated for each parameter from each model fit. Coverage probabilities are calculated from $\hat{\boldsymbol{\Omega}} \pm 1.96\text{SE}(\hat{\boldsymbol{\Omega}})$. The median [IQR] elapsed times for the approximate EM algorithm to converge and standard errors calculated was 7.422 [6.520, 8.624] seconds for $r = 5$, 5.237 [4.771, 5.801] seconds for $r = 10$ and 4.819 [4.380, 5.412] seconds for $r = 15$. *Total* computation time was 9.226 [8.373, 10.486] seconds for $r = 5$, 7.533 [7.076, 8.096] seconds for $r = 10$ and 7.727 [7.308, 8.389] seconds for $r = 15$.

B.2.3 Univariate joint models

Additional results from the simulations studies undertaken in Section 4.5.6. Tabulated results for the univariate joint models with true parameters given in Table 4.2 are presented in Tables B.15–B.18. The second set of results from the negative binomial with smaller fixed effects $\boldsymbol{\beta}$ is presented in Table B.19.

Parameter	Mean (SD)	SE	Bias	MSE	CP
$D_{1,1} = 0.200$	0.199 (0.016)	0.017	-0.001	0.000	0.96
$D_{2,2} = 0.050$	0.050 (0.004)	0.004	0.000	0.000	0.95
$\beta_0 = 0.000$	-0.007 (0.033)	0.032	-0.007	0.001	0.94
$\beta_1 = -0.100$	-0.101 (0.013)	0.012	-0.001	0.000	0.95
$\beta_2 = 0.100$	0.099 (0.021)	0.023	-0.001	0.000	0.95
$\beta_3 = -0.200$	-0.204 (0.049)	0.046	-0.004	0.002	0.91
$\sigma = 2.000$	2.001 (0.025)	0.025	0.001	0.001	0.96
$\gamma = 0.500$	0.488 (0.115)	0.108	-0.012	0.013	0.93
$\zeta = -0.200$	-0.200 (0.155)	0.168	0.000	0.024	0.97

Table B.15: Parameter estimates for univariate Gamma joint models. Median [IQR] elapsed time taken for approximate EM algorithm to converge and standard error calculation was 3.561 [3.405, 3.732] seconds and total computation time was 7.188 [6.870, 7.540] seconds.

Parameter	Mean (SD)	SE	Bias	MSE	CP
$D_{1,1} = 0.300$	0.297 (0.021)	0.021	-0.003	0.000	0.95
$\beta_0 = 1.000$	1.068 (0.033)	0.038	0.068	0.006	0.61
$\beta_1 = 0.050$	0.050 (0.004)	0.004	0.000	0.000	0.95
$\beta_2 = -0.050$	-0.047 (0.022)	0.026	0.003	0.000	0.98
$\beta_3 = 0.100$	0.086 (0.045)	0.052	-0.014	0.002	0.97
$\sigma = -0.300$	-0.307 (0.008)	0.008	-0.007	0.000	0.84
$\gamma = 0.500$	0.505 (0.170)	0.167	0.005	0.029	0.96
$\zeta = -0.200$	-0.213 (0.176)	0.171	-0.013	0.031	0.95

Table B.16: Parameter estimates for univariate underdispersed generalised Poisson joint models. Median [IQR] elapsed time taken for approximate EM algorithm to converge and standard error calculation was 4.437 [3.687, 5.891] seconds and total computation time was 6.832 [6.005, 8.402] seconds.

Parameter	Mean (SD)	SE	Bias	MSE	CP
$D_{1,1} = 0.700$	0.667 (0.052)	0.056	-0.033	0.004	0.90
$\beta_{10} = 0.500$	0.546 (0.060)	0.063	0.046	0.006	0.90
$\beta_{11} = -0.200$	-0.203 (0.012)	0.010	-0.003	0.000	0.92
$\beta_{12} = 0.050$	0.048 (0.038)	0.041	-0.002	0.001	0.98
$\beta_{13} = 0.400$	0.380 (0.080)	0.084	-0.020	0.007	0.95
$\sigma_{10} = 0.300$	0.251 (0.020)	0.019	-0.049	0.003	0.27
$\gamma_1 = 0.500$	0.548 (0.114)	0.108	0.048	0.015	0.91
$\zeta = -0.200$	-0.195 (0.170)	0.165	0.005	0.029	0.94

Table B.17: Parameter estimates for univariate overdispersed generalised Poisson joint models. Median [IQR] elapsed time taken for approximate EM algorithm to converge and standard error calculation was 4.022 [3.660, 4.375] seconds and total computation time was 5.739 [5.303, 6.232] seconds.

Parameter	Mean (SD)	SE	Bias	MSE	CP
$D_{1,1} = 0.500$	0.485 (0.047)	0.049	-0.015	0.002	0.94
$D_{2,2} = 0.090$	0.083 (0.009)	0.009	-0.007	0.000	0.86
$\beta_0 = 2.000$	2.086 (0.061)	0.049	0.086	0.011	0.56
$\beta_1 = -0.100$	-0.067 (0.021)	0.015	0.033	0.002	0.44
$\beta_2 = 0.100$	0.116 (0.041)	0.032	0.016	0.002	0.87
$\beta_3 = 0.200$	0.220 (0.082)	0.066	0.020	0.007	0.89
$\sigma = 1.000$	0.989 (0.037)	0.039	-0.011	0.001	0.95
$\gamma = 0.500$	0.521 (0.086)	0.083	0.021	0.008	0.92
$\zeta = -0.200$	-0.193 (0.173)	0.168	0.007	0.030	0.96

Table B.18: Parameter estimates for univariate negative binomial joint models. Median [IQR] elapsed time taken for approximate EM algorithm to converge and standard error calculation was 13.312 [12.507, 14.332] seconds and total computation time was 28.306 [27.379, 29.462] seconds.

Parameter	Mean (SD)	SE	Bias	MSE	CP
$D_{1,1} = 0.500$	0.447 (0.056)	0.058	-0.053	0.006	0.84
$D_{2,2} = 0.090$	0.078 (0.009)	0.010	-0.012	0.000	0.76
$\beta_{10} = 0.500$	0.481 (0.062)	0.054	-0.019	0.004	0.90
$\beta_{11} = 0.050$	0.040 (0.022)	0.017	-0.010	0.001	0.85
$\beta_{12} = 0.100$	0.111 (0.045)	0.035	0.011	0.002	0.85
$\beta_{13} = -0.100$	-0.109 (0.090)	0.070	-0.009	0.008	0.87
$\sigma_{10} = 1.000$	0.971 (0.066)	0.063	-0.029	0.005	0.92
$\gamma_1 = 0.500$	0.513 (0.095)	0.084	0.013	0.009	0.90
$\zeta = -0.200$	-0.206 (0.184)	0.174	-0.006	0.034	0.94

Table B.19: Parameter estimates for univariate negative binomial joint models with smaller $\boldsymbol{\beta}$ values. Median [IQR] elapsed time taken for approximate EM algorithm to converge and standard error calculation was 5.414 [4.567, 6.466] seconds and total computation time was 11.215 [10.346, 12.401] seconds.

B.2.4 Results from using the normal approximation in a Monte Carlo EM algorithm

	Gauss-Hermite quadrature					Antithetic Monte Carlo					Quasi Monte Carlo				
Parameter	Emp. Mean (SD)	Mean SE	Bias	MSE	CP	Emp. Mean (SD)	Mean SE	Bias	MSE	CP	Emp. Mean (SD)	Mean SE	Bias	MSE	CP
$D_{1,1} = 0.250$	0.248 (0.018)	0.021	-0.002	0.000	0.96	0.248 (0.018)	0.021	-0.002	0.000	0.97	0.249 (0.018)	0.021	-0.001	0.000	0.97
$D_{2,1} = 0.000$	0.000 (0.009)	0.009	0.000	0.000	0.93	0.000 (0.009)	0.009	0.000	0.000	0.93	0.000 (0.009)	0.009	0.000	0.000	0.93
$D_{3,1} = 0.125$	0.124 (0.018)	0.021	-0.001	0.000	0.96	0.129 (0.019)	0.021	0.004	0.000	0.96	0.128 (0.019)	0.021	0.003	0.000	0.96
$D_{4,1} = 0.000$	-0.002 (0.008)	0.009	-0.002	0.000	0.97	-0.002 (0.008)	0.009	-0.002	0.000	0.97	-0.002 (0.008)	0.009	-0.002	0.000	0.96
$D_{5,1} = 0.125$	0.115 (0.042)	0.046	-0.010	0.002	0.97	0.121 (0.044)	0.047	-0.004	0.002	0.97	0.120 (0.043)	0.046	-0.005	0.002	0.97
$D_{2,2} = 0.090$	0.089 (0.006)	0.007	-0.001	0.000	0.97	0.090 (0.006)	0.007	0.000	0.000	0.97	0.090 (0.006)	0.007	0.000	0.000	0.97
$D_{3,2} = 0.000$	0.001 (0.010)	0.012	0.001	0.000	0.97	0.001 (0.011)	0.012	0.001	0.000	0.98	0.001 (0.011)	0.012	0.001	0.000	0.98
$D_{4,2} = 0.000$	0.000 (0.005)	0.005	0.000	0.000	0.98	0.000 (0.005)	0.005	0.000	0.000	0.98	0.000 (0.005)	0.005	0.000	0.000	0.98
$D_{5,2} = 0.000$	0.000 (0.026)	0.028	0.000	0.001	0.98	0.000 (0.026)	0.028	0.000	0.001	0.98	0.000 (0.026)	0.028	0.000	0.001	0.98
$D_{3,3} = 0.500$	0.488 (0.032)	0.038	-0.012	0.001	0.95	0.516 (0.035)	0.041	0.016	0.001	0.99	0.513 (0.035)	0.041	0.013	0.001	0.99
$D_{4,3} = 0.000$	0.000 (0.012)	0.012	0.000	0.000	0.97	-0.002 (0.012)	0.012	-0.002	0.000	0.98	-0.001 (0.012)	0.012	-0.001	0.000	0.97
$D_{5,3} = 0.125$	0.128 (0.053)	0.061	0.003	0.003	0.98	0.154 (0.058)	0.064	0.029	0.004	0.96	0.148 (0.058)	0.063	0.023	0.004	0.95
$D_{4,4} = 0.090$	0.088 (0.007)	0.007	-0.002	0.000	0.95	0.087 (0.007)	0.007	-0.003	0.000	0.93	0.087 (0.007)	0.007	-0.003	0.000	0.92
$D_{5,4} = 0.000$	0.001 (0.024)	0.028	0.001	0.001	0.99	0.001 (0.024)	0.028	0.001	0.001	0.98	0.001 (0.024)	0.027	0.001	0.001	0.99
$D_{5,5} = 2.000$	1.674 (0.238)	0.258	-0.326	0.162	0.71	1.685 (0.245)	0.259	-0.315	0.159	0.73	1.650 (0.241)	0.254	-0.350	0.180	0.63
$\beta_{10} = -2.000$	-2.007 (0.040)	0.037	-0.007	0.002	0.91	-2.026 (0.040)	0.037	-0.026	0.002	0.86	-2.024 (0.040)	0.037	-0.024	0.002	0.88
$\beta_{11} = 0.100$	0.095 (0.015)	0.016	-0.005	0.000	0.95	0.100 (0.016)	0.016	0.000	0.000	0.94	0.103 (0.017)	0.016	0.003	0.000	0.92
$\beta_{11} = -0.100$	-0.102 (0.026)	0.026	-0.002	0.001	0.94	-0.100 (0.026)	0.026	0.000	0.001	0.94	-0.100 (0.026)	0.026	0.000	0.001	0.93
$\beta_{11} = 0.200$	0.199 (0.053)	0.052	-0.001	0.003	0.96	0.200 (0.053)	0.051	0.000	0.003	0.95	0.200 (0.053)	0.052	0.000	0.003	0.95
$\beta_{20} = 2.000$	2.003 (0.043)	0.049	0.003	0.002	0.97	1.878 (0.056)	0.052	-0.122	0.018	0.40	1.889 (0.055)	0.052	-0.111	0.015	0.49
$\beta_{21} = -0.100$	-0.095 (0.015)	0.016	0.005	0.000	0.91	-0.090 (0.017)	0.018	0.010	0.000	0.90	-0.091 (0.016)	0.017	0.009	0.000	0.91
$\beta_{21} = 0.100$	0.101 (0.034)	0.035	0.001	0.001	0.96	0.110 (0.037)	0.033	0.010	0.001	0.94	0.109 (0.037)	0.034	0.009	0.001	0.94
$\beta_{21} = 0.200$	0.196 (0.065)	0.069	-0.004	0.004	0.95	0.211 (0.070)	0.067	0.011	0.005	0.93	0.210 (0.069)	0.067	0.010	0.005	0.94
$\beta_{30} = 1.000$	0.889 (0.139)	0.123	-0.111	0.032	0.79	0.860 (0.141)	0.123	-0.140	0.039	0.72	0.870 (0.138)	0.122	-0.130	0.036	0.77
$\beta_{31} = -1.000$	-0.991 (0.042)	0.047	0.009	0.002	0.96	-0.991 (0.042)	0.047	0.009	0.002	0.96	-0.983 (0.042)	0.046	0.017	0.002	0.95
$\beta_{31} = 1.000$	1.034 (0.074)	0.085	0.034	0.007	0.98	1.037 (0.075)	0.085	0.037	0.007	0.97	1.026 (0.075)	0.084	0.026	0.006	0.99
$\beta_{31} = -1.000$	-1.035 (0.179)	0.155	-0.035	0.033	0.91	-1.034 (0.180)	0.155	-0.034	0.033	0.91	-1.024 (0.178)	0.154	-0.024	0.032	0.92
$\sigma_1^2 = 0.160$	0.160 (0.004)	0.004	0.000	0.000	0.97	0.160 (0.004)	0.004	0.000	0.000	0.96	0.159 (0.004)	0.004	-0.001	0.000	0.96
$\gamma_1 = 0.500$	0.508 (0.089)	0.095	0.008	0.008	0.97	0.510 (0.090)	0.096	0.010	0.008	0.97	0.512 (0.091)	0.097	0.012	0.008	0.98
$\gamma_2 = -0.500$	-0.482 (0.085)	0.087	0.018	0.007	0.95	-0.485 (0.085)	0.089	0.015	0.007	0.96	-0.490 (0.086)	0.090	0.010	0.007	0.96
$\gamma_3 = 0.500$	0.462 (0.094)	0.095	-0.038	0.010	0.96	0.464 (0.094)	0.096	-0.036	0.010	0.98	0.496 (0.101)	0.098	-0.004	0.010	0.96
$\zeta = -0.200$	-0.213 (0.174)	0.182	-0.013	0.030	0.96	-0.219 (0.175)	0.184	-0.019	0.031	0.96	-0.220 (0.176)	0.185	-0.020	0.031	0.96

Table B.20: Comparisons between joint models fit with different E-steps approaches, each employing the normal approximation in Section 3.2.1. 'Emp. Mean (SD)' denotes the average estimated value with the standard deviation of parameter estimates and Mean SE the mean standard error calculated for each parameter from each model fit. Coverage probabilities are calculated from $\hat{\boldsymbol{\Omega}} \pm 1.96\text{SE}(\hat{\boldsymbol{\Omega}})$. Joint models fit by Gauss-Hermite quadrature had median [IQR] elapsed time for model convergence and calculation standard errors 9.03 [8.43, 9.56] seconds and total computation time 12.05 [11.40, 12.65], with median number of iterations 7. Joint models fit by Antithetic Monte Carlo had median [IQR] elapsed time for model convergence and calculation of standard errors 24.42 [17.49, 34.69] seconds and total computation time 27.98 [20.58, 37.78], with median number of iterations 14. Joint models fit by Quasi Monte Carlo via the Sobol sequence had median [IQR] elapsed time for model convergence and calculation of standard errors 25.93 [16.36, 44.84] seconds and total computation time 29.59 [19.74, 48.17], with median number of iterations 14.

B.3 Additional results from Chapter 7

B.3.1 Survival sub-model selection

The 'full' four-variate model mentioned in Section 7.3.1; covariates were (in order): `drug`; `age`; `sex`; `histologic`, fit to the Cox PH model

$$\lambda_i(t) = \lambda_0(t) \exp\{\texttt{age} \times \zeta_1 + \texttt{drug} \times \zeta_2 + \texttt{sex} \times \zeta_3 + \texttt{histologic} \times \zeta_4\} \tag{B.2}$$

which had resultant parameter estimates

Parameter	Estimate	exp{Estimate}	Standard Error	$\lvert Z \rvert$	p-value
ζ_1	0.42	1.52	0.09	4.55	< 0.001
ζ_2	-0.09	0.91	0.17	-0.52	0.605
ζ_3	-0.44	0.64	0.22	-1.97	0.049
ζ_4	0.99	2.68	0.23	4.20	< 0.001

Table B.21: Parameter estimates, presented with their standard errors and exponentiated value for the 'full' four-variate Cox PH model fit to PBC data.

B.3.2 Intermediate multivariate joint model fits

In Section 7.4.2 we consider three intermediary multivariate joint models housing the PBC biomarkers in three loose groups of liver function. Here we present the written-out joint models along with full parameter estimates of the fitted models and the variance-covariance matrices with correlations between random effects additionally shown.

The bivariate joint model we fit to '*Blood clotting and flow*' was

$$\begin{cases} \begin{cases} \log(\mathbb{E}[\text{Prothrombin time}|\boldsymbol{b}_{i1}]) & = (\beta_{10} + b_{i10}) + (\beta_{11} + b_{i11})\, t \\ & \quad + \texttt{H}_i \times \beta_{12} + \texttt{S}_i \times \beta_{13} \\ \log(\boldsymbol{\varphi}_{i1}) & = \sigma_{10} + \sigma_{11} t \end{cases} \\ \begin{cases} \log(\mathbb{E}[\text{Platelet count}|\boldsymbol{b}_{i2}]) & = (\beta_{20} + b_{i20}) + (\beta_{21} + b_{i21})\, N_1(t) \\ & \quad + (\beta_{22} + b_{i22})\, N_2(t) + (\beta_{23} + b_{i23})\, N_3(t) \\ & \quad + \texttt{A}_i \times \beta_{24} + \texttt{H}_i \times \beta_{25} \\ \boldsymbol{\varphi}_{i2} & = \sigma_{20} + \sigma_{21} t \end{cases} \\ \lambda_i(t) = \lambda_0(t) \exp\Big\{\texttt{A}_i \times \zeta_1 + \texttt{S}_i \times \zeta_2 + \texttt{H}_i \times \zeta_3 + \sum_{k=1}^{2} \gamma_k \boldsymbol{W}_k(t)^\top \boldsymbol{b}_{ik}\Big\}, \end{cases} \tag{B.3}$$

where subject i's age, sex and histologic state is given by $\texttt{A}_i$, $\texttt{S}_i$ and $\texttt{H}_i$ respectively and $N_1(t), \ldots, N_3(t)$ denotes the set of natural cubic splines with knots at tertiles of follow-up.

Left braces are used to separate each sub-model (and its dispersion model) visually, with the left-most brace denoting that these are to be jointly modelled.

In a similar fashion we elucidate the model for '*Liver enzymes*'

$$\begin{cases} \begin{cases} \log(\mathbb{E}[\text{Alkaline}|\boldsymbol{b}_{i1}]) & = (\beta_{10} + b_{i10}) + (\beta_{11} + b_{i11})\, N_1(t) \\ & \quad + (\beta_{12} + b_{i12})\, N_2(t) + (\beta_{13} + b_{i13})\, N_3(t) \\ & \quad + \mathtt{A}_i \times \beta_{14} + \mathtt{H}_i \times \beta_{15} \\ \log(\boldsymbol{\varphi}_{i1}) & = \sigma_{10} + \sigma_{11}\mathtt{H}_i \end{cases} \\ \begin{cases} \log(\text{AST}) & = (\beta_{20} + b_{i20}) + (\beta_{21} + b_{i21})\, t + (\beta_{22} + b_{i22})\, t^2 \\ & \quad + \mathtt{A}_i \times \beta_{23} + \mathtt{H}_i \times \beta_{24} + \varepsilon_{i2}(t) \end{cases} \\ \begin{cases} \log(\text{Serum bilirubin}) & = (\beta_{30} + b_{i30}) + (\beta_{31} + b_{i31})\, N_1(t) \\ & \quad + (\beta_{32} + b_{i32})\, N_2(t) + (\beta_{33} + b_{i33})\, N_3(t) \\ & \quad + \mathtt{H}_i \times \beta_{34} + \mathtt{S}_i \times \beta_{35} + \varepsilon_{i3}(t) \end{cases} \\ \lambda_i(t) = \lambda_0(t) \exp\left\{ \mathtt{A}_i \times \zeta_1 + \mathtt{S}_i \times \zeta_2 + \mathtt{H}_i \times \zeta_3 + \sum_{k=1}^{3} \gamma_k \boldsymbol{W}_k(t)^\top \boldsymbol{b}_{ik} \right\}, \end{cases} \tag{B.4}$$

and finally '*Liver health and function*'

$$\begin{cases} \begin{cases} \text{logit}(\mathbb{E}[\text{Hepatomegaly}|\boldsymbol{b}_{i1}]) & = (\beta_{10} + b_{i10}) + (\beta_{11} + b_{i11})\, t \\ & \quad + \mathtt{H}_i \times \beta_{12} + \mathtt{S}_i \times \beta_{13} \end{cases} \\ \left\{ \text{Albumin} \quad = (\beta_{20} + b_{i20}) + (\beta_{21} + b_{i21})\, t + \mathtt{A}_i \times \beta_{22} + \mathtt{H}_i \times \beta_{23} + \varepsilon_{i2}(t) \right. \\ \lambda_i(t) = \lambda_0(t) \exp\left\{ \mathtt{A}_i \times \zeta_1 + \mathtt{S}_i \times \zeta_2 + \mathtt{H}_i \times \zeta_3 + \sum_{k=1}^{2} \gamma_k \boldsymbol{W}_k(t)^\top \boldsymbol{b}_{ik} \right\}. \end{cases} \tag{B.5}$$

The full parameter estimates are given in Tables B.22–B.24 and correspond to (B.3)–(B.5).

Parameter	Estimate (SE)	95% CI	Parameter	Estimate (SE)	95% CI
β_{10}	2.339 (0.018)	[2.304, 2.375]	β_{23}	-0.627 (0.122)	[-0.866, -0.387]
β_{11}	0.016 (0.001)	[0.013, 0.018]	β_{24}	-0.031 (0.021)	[-0.072, 0.010]
β_{12}	0.047 (0.015)	[0.018, 0.077]	β_{25}	-0.212 (0.054)	[-0.319, -0.105]
β_{13}	-0.017 (0.015)	[-0.047, 0.012]	σ_{20}	1.111 (0.033)	[1.046, 1.176]
σ_{10}	4.901 (0.033)	[4.835, 4.967]	σ_{21}	-0.015 (0.008)	[-0.031, 0.000]
σ_{11}	0.121 (0.011)	[0.010, 0.142]	γ_2	0.078 (0.268)	[-0.447, 0.603]
γ_1	13.026 (1.615)	[9.860, 16.192]	ζ_1	0.393 (0.107)	[0.184, 0.603]
β_{20}	5.666 (0.048)	[5.571, 5.761]	ζ_2	-0.232 (0.297)	[-0.813, 0.349]
β_{21}	-0.313 (0.063)	[-0.437, -0.189]	ζ_3	1.097 (0.320)	[0.471, 1.724]
β_{22}	-0.674 (0.077)	[-0.825, -0.522]			

Table B.22: Parameter estimates (SE: standard error) for the '*Blood clotting and flow*' bivariate joint model (B.3). The elapsed time for the approximate EM algorithm to converge and SE calculation was 30.511 seconds. Total computation time was 34.564 seconds.

Parameter	Estimate (SE)	95% CI	Parameter	Estimate (SE)	95% CI
β_{10}	7.428 (0.098)	[7.235, 7.621]	σ_2^2	0.070 (0.001)	[0.067, 0.072]
β_{11}	-0.284 (0.149)	[-0.575, 0.008]	γ_2	0.134 (0.570)	[-0.982, 1.250]
β_{12}	-0.822 (0.203)	[-1.220, -0.424]	β_{30}	0.158 (0.260)	[-0.351, 0.668]
β_{13}	-0.736 (0.347)	[-1.416, -0.056]	β_{31}	1.044 (0.166)	[0.718, 1.369]
β_{14}	-0.129 (0.038)	[-0.205, -0.054]	β_{32}	1.396 (0.219)	[0.966, 1.826]
β_{15}	0.096 (0.094)	[-0.087, 0.280]	β_{33}	1.270 (0.394)	[0.497, 2.042]
σ_{10}	2.369 (0.046)	[2.278, 2.459]	β_{34}	0.590 (0.161)	[0.274, 0.907]
σ_{11}	0.530 (0.055)	[0.422, 0.639]	β_{35}	-0.028 (0.236)	[-0.491, 0.435]
γ_1	-0.651 (0.357)	[-1.350, 0.049]	σ_3^2	0.076 (0.003)	[0.070, 0.082]
β_{20}	4.651 (0.067)	[4.519, 4.782]	γ_3	1.478 (0.180)	[1.125, 1.831]
β_{21}	0.023 (0.018)	[-0.012, 0.057]	ζ_1	0.672 (0.104)	[0.467, 0.876]
β_{22}	-0.002 (0.002)	[-0.006, 0.003]	ζ_2	0.177 (0.499)	[-0.801, 1.155]
β_{23}	-0.120 (0.024)	[-0.166, -0.073]	ζ_3	1.728 (0.386)	[0.971, 2.485]
β_{24}	0.158 (0.068)	[0.024, 0.292]			

Table B.23: Parameter estimates (SE: standard error) for the '*Liver enzymes*' trivariate joint model (B.4). The elapsed time for the approximate EM algorithm to converge and SE calculation was 109.986 seconds. Total computation time was 118.019 seconds.

Parameter	Estimate (SE)	95% CI	Parameter	Estimate (SE)	95% CI
β_{10}	-1.293 (0.393)	[-2.062, -0.523]	β_{22}	-0.071 (0.021)	[-0.112, -0.030]
β_{11}	0.196 (0.038)	[0.121, 0.270]	β_{23}	-0.329 (0.057)	[-0.440, -0.218]
β_{12}	3.272 (0.356)	[2.575, 3.969]	σ_2^2	0.098 (0.002)	[0.095, 0.101]
β_{13}	-1.337 (0.342)	[-2.007, -0.667]	γ_2	-2.882 (0.447)	[-3.759, -2.005]
γ_1	0.114 (0.092)	[-0.066, 0.294]	ζ_1	0.468 (0.114)	[0.245, 0.691]
β_{20}	3.795 (0.052)	[3.693, 3.898]	ζ_2	-0.628 (0.268)	[-1.154, -0.101]
β_{21}	-0.084 (0.005)	[-0.094, -0.074]	ζ_3	1.713 (0.330)	[1.066, 2.360]

Table B.24: Parameter estimates (SE: standard error) for the '*Liver health and function*' bivariate joint model (B.5). The elapsed time for the approximate EM algorithm to converge and SE calculation was 21.170 seconds. Total computation time was 23.004 seconds.

Finally, the variance-covariance matrices are presented in Tables B.25–B.27, following the same order as before.

	$D_{1,0}$	$D_{1,1}$	$D_{2,0}$	$D_{2,1}$	$D_{2,2}$	$D_{2,3}$
$D_{1,0}$	0.0054	-0.1975	-0.3761	-0.2160	-0.1112	-0.0061
$D_{1,1}$	-0.0001	0.0001	-0.0859	-0.0609	-0.3930	-0.3347
$D_{2,0}$	-0.0094	-0.0003	0.1155	-0.2034	-0.2229	-0.2462
$D_{2,1}$	-0.0075	-0.0003	-0.0328	0.2250	0.4991	0.2689
$D_{2,2}$	-0.0050	-0.0021	-0.0464	0.1450	0.3752	0.5238
$D_{2,3}$	-0.0003	-0.0018	-0.0494	0.0752	0.1893	0.3480

Table B.25: Covariance matrix estimates for (B.3); resulting correlation estimates are presented on the upper triangle of the matrix. The diagonal elements are shaded light grey to facilitate easier reading. Row and column names are given as $D_{k,e}$ where k denotes the longitudinal response and e the random effect index (i.e. 0: intercept).

	$D_{1,0}$	$D_{1,1}$	$D_{1,2}$	$D_{1,3}$	$D_{2,0}$	$D_{2,1}$	$D_{2,2}$	$D_{3,0}$	$D_{3,1}$	$D_{3,2}$	$D_{3,3}$
$D_{1,0}$	0.4610	-0.4297	-0.7110	-0.5690	0.4826	0.3501	-0.3682	0.3362	0.1716	0.2273	0.1876
$D_{1,1}$	-0.1929	0.4370	0.4715	0.1910	-0.4529	0.3303	-0.1331	-0.3244	0.2790	0.3547	0.1674
$D_{1,2}$	-0.4623	0.2985	0.9170	0.3469	-0.1451	0.0384	-0.0186	-0.0026	-0.0249	0.0746	-0.0107
$D_{1,3}$	-0.2658	0.0869	0.2286	0.4734	-0.3492	-0.5582	0.7227	-0.1897	-0.3379	-0.1198	0.2631
$D_{2,0}$	0.1306	-0.1194	-0.0554	-0.0958	0.1590	0.1065	-0.1790	0.6259	0.0261	0.2148	0.1628
$D_{2,1}$	0.0275	0.0253	0.0043	-0.0444	0.0049	0.0134	-0.9087	0.2627	0.7350	0.7012	0.1527
$D_{2,2}$	-0.0023	-0.0008	-0.0002	0.0047	-0.0007	-0.0010	0.0001	-0.2680	-0.5710	-0.4651	0.2022
$D_{3,0}$	0.2171	-0.2040	-0.0024	-0.1242	0.2374	0.0289	-0.0024	0.9049	0.0312	0.2127	0.2067
$D_{3,1}$	0.1495	0.2367	-0.0305	-0.2984	0.0134	0.1091	-0.0069	0.0380	1.6469	0.7962	0.2695
$D_{3,2}$	0.2192	0.3330	0.1015	-0.1170	0.1216	0.1152	-0.0062	0.2873	1.4512	2.0169	0.5782
$D_{3,3}$	0.1767	0.1535	-0.0142	0.2512	0.0901	0.0245	0.0026	0.2728	0.4798	1.1394	1.9251

Table B.26: Covariance matrix estimates for (B.4); resulting correlation estimates are presented on the upper triangle of the matrix. The diagonal elements are shaded light grey to facilitate easier reading. Row and column names are given as $D_{k,e}$ where k denotes the longitudinal response and e the random effect index (i.e. 0: intercept).

	$D_{1,0}$	$D_{1,1}$	$D_{2,0}$	$D_{2,1}$
$D_{1,0}$	4.0678	-0.2759	-0.5730	-0.1204
$D_{1,1}$	-0.2333	0.1758	0.0438	-0.5481
$D_{2,0}$	-0.3865	0.0061	0.1119	-0.0709
$D_{2,1}$	-0.0123	-0.0116	-0.0012	0.0026

Table B.27: Covariance matrix estimates for (B.5); resulting correlation estimates are presented on the upper triangle of the matrix. The diagonal elements are shaded light grey to facilitate easier reading. Row and column names are given as $D_{k,e}$ where k denotes the longitudinal response and e the random effect index (i.e. 0: intercept).

B.3.3 Classification metrics for the final PBC model

In Tables B.28–B.30 we present the classification metrics which were outlined in Tables 6.1 and 6.2, as well as the F_1 score and Youden's J_Y statistic (6.13) evaluated at different probabilistic thresholds $\boldsymbol{c} = (1.00, 0.99, \ldots, 0.01, 0.00)^\top$.

c_j	TP	TN	FP	FN	TPR	FPR	PPV	NPV	Acc	F_1	J_Y
1.00	33	0	231	0	1.00	1.00	0.12	0.00	0.12	0.22	0.00
0.99	32	67	164	1	0.97	0.71	0.16	0.99	0.38	0.28	0.26
0.98	32	110	121	1	0.97	0.52	0.21	0.99	0.54	0.34	0.45
0.97	32	139	92	1	0.97	0.40	0.26	0.99	0.65	0.41	0.57
0.96	32	151	80	1	0.97	0.35	0.29	0.99	0.69	0.44	0.62
0.95	32	159	72	1	0.97	0.31	0.31	0.99	0.72	0.47	0.66
0.94	32	166	65	1	0.97	0.28	0.33	0.99	0.75	0.49	0.69
0.93	32	176	55	1	0.97	0.24	0.37	0.99	0.79	0.53	0.73
0.92	31	179	52	2	0.94	0.23	0.37	0.99	0.80	0.53	0.71
0.91	31	184	47	2	0.94	0.20	0.40	0.99	0.81	0.56	0.74
0.90	30	185	46	3	0.91	0.20	0.39	0.98	0.81	0.55	0.71
0.89	29	191	40	4	0.88	0.17	0.42	0.98	0.83	0.57	0.71
0.88	29	193	38	4	0.88	0.16	0.43	0.98	0.84	0.58	0.71
0.87	29	197	34	4	0.88	0.15	0.46	0.98	0.86	0.60	0.73
0.86	28	198	33	5	0.85	0.14	0.46	0.98	0.86	0.60	0.71
0.85	27	198	33	6	0.82	0.14	0.45	0.97	0.85	0.58	0.68
0.84	27	199	32	6	0.82	0.14	0.46	0.97	0.86	0.59	0.68
0.83	27	200	31	6	0.82	0.13	0.47	0.97	0.86	0.59	0.68
0.81	27	202	29	6	0.82	0.13	0.48	0.97	0.87	0.61	0.69
0.80	27	204	27	6	0.82	0.12	0.50	0.97	0.88	0.62	0.70
0.77	27	207	24	6	0.82	0.10	0.53	0.97	0.89	0.64	0.71
0.76	27	211	20	6	0.82	0.09	0.57	0.97	0.90	0.67	0.73
0.75	26	212	19	7	0.79	0.08	0.58	0.97	0.90	0.67	0.71
0.74	26	213	18	7	0.79	0.08	0.59	0.97	0.91	0.68	0.71
0.73	26	214	17	7	0.79	0.07	0.60	0.97	0.91	0.68	0.71
0.71	25	214	17	8	0.76	0.07	0.60	0.96	0.91	0.67	0.68
0.70	24	215	16	9	0.73	0.07	0.60	0.96	0.91	0.66	0.66
0.68	23	215	16	10	0.70	0.07	0.59	0.96	0.90	0.64	0.63
0.67	22	215	16	11	0.67	0.07	0.58	0.95	0.90	0.62	0.60
0.65	20	215	16	13	0.61	0.07	0.56	0.94	0.89	0.58	0.54
0.64	19	216	15	14	0.58	0.06	0.56	0.94	0.89	0.57	0.51
0.61	19	217	14	14	0.58	0.06	0.58	0.94	0.89	0.58	0.52
0.60	19	219	12	14	0.58	0.05	0.61	0.94	0.90	0.59	0.52
0.59	19	220	11	14	0.58	0.05	0.63	0.94	0.91	0.60	0.53
0.57	17	222	9	16	0.52	0.04	0.65	0.93	0.91	0.58	0.48
0.55	16	222	9	17	0.48	0.04	0.64	0.93	0.90	0.55	0.45
0.53	15	223	8	18	0.45	0.03	0.65	0.93	0.90	0.54	0.42
0.47	15	224	7	18	0.45	0.03	0.68	0.93	0.91	0.55	0.42
0.43	14	224	7	19	0.42	0.03	0.67	0.92	0.90	0.52	0.39
0.37	13	225	6	20	0.39	0.03	0.68	0.92	0.90	0.50	0.37
0.36	13	226	5	20	0.39	0.02	0.72	0.92	0.91	0.51	0.37
0.34	12	226	5	21	0.36	0.02	0.71	0.91	0.90	0.48	0.34
0.31	11	226	5	22	0.33	0.02	0.69	0.91	0.90	0.45	0.31
0.30	10	226	5	23	0.30	0.02	0.67	0.91	0.89	0.42	0.28
0.29	9	228	3	24	0.27	0.01	0.75	0.90	0.90	0.40	0.26
0.28	8	229	2	25	0.24	0.01	0.80	0.90	0.90	0.37	0.23
0.24	7	229	2	26	0.21	0.01	0.78	0.90	0.89	0.33	0.20
0.18	6	229	2	27	0.18	0.01	0.75	0.89	0.89	0.29	0.17
0.15	4	229	2	29	0.12	0.01	0.67	0.89	0.88	0.21	0.11
0.12	3	230	1	30	0.09	0.00	0.75	0.88	0.88	0.16	0.09
0.09	3	231	0	30	0.09	0.00	1.00	0.89	0.89	0.17	0.09
0.08	2	231	0	31	0.06	0.00	1.00	0.88	0.88	0.11	0.06
0.06	1	231	0	32	0.03	0.00	1.00	0.88	0.88	0.06	0.03
0.00	0	231	0	33	0.00	0.00	0.00	0.87	0.88	0.00	0.00

Table B.28: Window: [2.0, 3.5]. 'Acc': Accuracy. Probabilistic thresholds c_j which are repeated are omitted and the greater presented.

c_j	TP	TN	FP	FN	TPR	FPR	PPV	NPV	Acc	F_1	J_Y
1.00	40	0	184	0	1.00	1.00	0.18	0.00	0.18	0.30	0.00
0.99	40	26	158	0	1.00	0.86	0.20	1.00	0.29	0.34	0.14
0.98	40	47	137	0	1.00	0.74	0.23	1.00	0.39	0.37	0.26
0.97	40	68	116	0	1.00	0.63	0.26	1.00	0.48	0.41	0.37
0.96	37	83	101	3	0.93	0.55	0.27	0.97	0.54	0.42	0.38
0.95	36	94	90	4	0.90	0.49	0.29	0.96	0.58	0.43	0.41
0.94	35	100	84	5	0.88	0.46	0.29	0.95	0.60	0.44	0.42
0.93	35	110	74	5	0.88	0.40	0.32	0.96	0.65	0.47	0.47
0.92	35	112	72	5	0.88	0.39	0.33	0.96	0.66	0.48	0.48
0.91	34	116	68	6	0.85	0.37	0.33	0.95	0.67	0.48	0.48
0.90	34	121	63	6	0.85	0.34	0.35	0.95	0.69	0.50	0.51
0.89	33	123	61	7	0.82	0.33	0.35	0.95	0.70	0.49	0.49
0.88	33	126	58	7	0.82	0.32	0.36	0.95	0.71	0.50	0.51
0.87	33	132	52	7	0.82	0.28	0.39	0.95	0.74	0.53	0.54
0.86	33	134	50	7	0.82	0.27	0.40	0.95	0.75	0.54	0.55
0.85	32	135	49	8	0.80	0.27	0.40	0.94	0.75	0.53	0.53
0.84	32	140	44	8	0.80	0.24	0.42	0.95	0.77	0.55	0.56
0.83	31	143	41	9	0.78	0.22	0.43	0.94	0.78	0.55	0.55
0.82	31	144	40	9	0.78	0.22	0.44	0.94	0.78	0.56	0.56
0.81	30	146	38	10	0.75	0.21	0.44	0.94	0.79	0.56	0.54
0.78	29	148	36	11	0.72	0.20	0.45	0.93	0.79	0.55	0.53
0.77	29	150	34	11	0.72	0.18	0.46	0.93	0.80	0.56	0.54
0.76	29	152	32	11	0.72	0.17	0.48	0.93	0.81	0.57	0.55
0.74	28	152	32	12	0.70	0.17	0.47	0.93	0.80	0.56	0.53
0.73	28	153	31	12	0.70	0.17	0.47	0.93	0.81	0.57	0.53
0.72	28	156	28	12	0.70	0.15	0.50	0.93	0.82	0.58	0.55
0.70	27	157	27	13	0.68	0.15	0.50	0.92	0.82	0.57	0.53
0.67	26	158	26	14	0.65	0.14	0.50	0.92	0.82	0.57	0.51
0.64	25	159	25	15	0.62	0.14	0.50	0.91	0.82	0.56	0.49
0.63	25	160	24	15	0.62	0.13	0.51	0.91	0.83	0.56	0.49
0.61	24	161	23	16	0.60	0.12	0.51	0.91	0.83	0.55	0.48
0.57	24	162	22	16	0.60	0.12	0.52	0.91	0.83	0.56	0.48
0.56	23	163	21	17	0.57	0.11	0.52	0.91	0.83	0.55	0.46
0.55	23	165	19	17	0.57	0.10	0.55	0.91	0.84	0.56	0.47
0.52	21	166	18	19	0.53	0.10	0.54	0.90	0.83	0.53	0.43
0.50	21	167	17	19	0.53	0.09	0.55	0.90	0.84	0.54	0.43
0.48	21	169	15	19	0.53	0.08	0.58	0.90	0.85	0.55	0.44
0.44	20	170	14	20	0.50	0.08	0.59	0.89	0.85	0.54	0.42
0.41	19	170	14	21	0.47	0.08	0.58	0.89	0.84	0.52	0.40
0.40	19	171	13	21	0.47	0.07	0.59	0.89	0.85	0.53	0.40
0.39	18	171	13	22	0.45	0.07	0.58	0.89	0.84	0.51	0.38
0.36	18	172	12	22	0.45	0.07	0.60	0.89	0.85	0.51	0.38
0.33	18	173	11	22	0.45	0.06	0.62	0.89	0.85	0.52	0.39
0.31	17	173	11	23	0.42	0.06	0.61	0.88	0.85	0.50	0.37
0.29	17	174	10	23	0.42	0.05	0.63	0.88	0.85	0.51	0.37
0.28	16	174	10	24	0.40	0.05	0.62	0.88	0.85	0.48	0.35
0.26	15	174	10	25	0.38	0.05	0.60	0.87	0.84	0.46	0.32
0.25	15	176	8	25	0.38	0.04	0.65	0.88	0.85	0.48	0.33
0.24	14	176	8	26	0.35	0.04	0.64	0.87	0.85	0.45	0.31
0.21	14	177	7	26	0.35	0.04	0.67	0.87	0.85	0.46	0.31
0.20	13	177	7	27	0.33	0.04	0.65	0.87	0.85	0.43	0.29
0.17	12	177	7	28	0.30	0.04	0.63	0.86	0.84	0.41	0.26
0.15	11	177	7	29	0.28	0.04	0.61	0.86	0.84	0.38	0.24
0.12	10	178	6	30	0.25	0.03	0.62	0.86	0.84	0.36	0.22
0.11	7	178	6	33	0.17	0.03	0.54	0.84	0.83	0.26	0.14
0.09	6	179	5	34	0.15	0.03	0.55	0.84	0.83	0.24	0.12
0.08	6	180	4	34	0.15	0.02	0.60	0.84	0.83	0.24	0.13
0.04	5	181	3	35	0.12	0.02	0.62	0.84	0.83	0.21	0.11
0.02	4	182	2	36	0.10	0.01	0.67	0.83	0.83	0.17	0.09
0.01	2	182	2	38	0.05	0.01	0.50	0.83	0.82	0.09	0.04
0.00	0	184	0	40	0.00	0.00	0.00	0.82	0.82	0.00	0.00

Table B.29: Window: [3.5, 7.0]. 'Acc': Accuracy. Probabilistic thresholds c_j which are repeated are ommitted and the greater presented.

c_j	TP	TN	FP	FN	TPR	FPR	PPV	NPV	Acc	F_1	J_Y
1.00	29	0	93	0	1.00	1.00	0.24	0.00	0.24	0.38	0.00
0.99	29	1	92	0	1.00	0.99	0.24	1.00	0.25	0.39	0.01
0.98	28	4	89	1	0.97	0.96	0.24	0.80	0.26	0.38	0.01
0.97	28	8	85	1	0.97	0.91	0.25	0.89	0.30	0.39	0.05
0.96	28	14	79	1	0.97	0.85	0.26	0.93	0.34	0.41	0.12
0.95	28	22	71	1	0.97	0.76	0.28	0.96	0.41	0.44	0.20
0.94	28	28	65	1	0.97	0.70	0.30	0.97	0.46	0.46	0.27
0.93	27	29	64	2	0.93	0.69	0.30	0.94	0.46	0.45	0.24
0.92	26	31	62	3	0.90	0.67	0.30	0.91	0.47	0.44	0.23
0.91	26	38	55	3	0.90	0.59	0.32	0.93	0.52	0.47	0.31
0.90	26	42	51	3	0.90	0.55	0.34	0.93	0.56	0.49	0.35
0.89	26	44	49	3	0.90	0.53	0.35	0.94	0.57	0.50	0.37
0.88	26	46	47	3	0.90	0.51	0.36	0.94	0.59	0.51	0.39
0.87	25	46	47	4	0.86	0.51	0.35	0.92	0.58	0.50	0.36
0.86	25	49	44	4	0.86	0.47	0.36	0.92	0.61	0.51	0.39
0.85	25	52	41	4	0.86	0.44	0.38	0.93	0.63	0.53	0.42
0.84	25	53	40	4	0.86	0.43	0.38	0.93	0.64	0.53	0.43
0.83	25	56	37	4	0.86	0.40	0.40	0.93	0.66	0.55	0.46
0.82	25	57	36	4	0.86	0.39	0.41	0.93	0.67	0.56	0.47
0.81	25	62	31	4	0.86	0.33	0.45	0.94	0.71	0.59	0.53
0.80	25	63	30	4	0.86	0.32	0.45	0.94	0.72	0.60	0.54
0.78	25	65	28	4	0.86	0.30	0.47	0.94	0.74	0.61	0.56
0.77	25	66	27	4	0.86	0.29	0.48	0.94	0.75	0.62	0.57
0.76	24	66	27	5	0.83	0.29	0.47	0.93	0.74	0.60	0.54
0.74	24	68	25	5	0.83	0.27	0.49	0.93	0.75	0.62	0.56
0.73	24	70	23	5	0.83	0.25	0.51	0.93	0.77	0.63	0.58
0.72	24	71	22	5	0.83	0.24	0.52	0.93	0.78	0.64	0.59
0.71	24	72	21	5	0.83	0.23	0.53	0.94	0.79	0.65	0.60
0.70	23	72	21	6	0.79	0.23	0.52	0.92	0.78	0.63	0.57
0.68	23	73	20	6	0.79	0.22	0.53	0.92	0.79	0.64	0.58
0.64	23	74	19	6	0.79	0.20	0.55	0.92	0.80	0.65	0.59
0.63	23	75	18	6	0.79	0.19	0.56	0.93	0.80	0.66	0.60
0.62	22	75	18	7	0.76	0.19	0.55	0.91	0.80	0.64	0.57
0.57	22	76	17	7	0.76	0.18	0.56	0.92	0.80	0.65	0.58
0.53	21	76	17	8	0.72	0.18	0.55	0.90	0.80	0.63	0.54
0.48	20	77	16	9	0.69	0.17	0.56	0.90	0.80	0.62	0.52
0.47	19	77	16	10	0.66	0.17	0.54	0.89	0.79	0.59	0.48
0.46	18	77	16	11	0.62	0.17	0.53	0.87	0.78	0.57	0.45
0.44	18	78	15	11	0.62	0.16	0.55	0.88	0.79	0.58	0.46
0.39	18	79	14	11	0.62	0.15	0.56	0.88	0.80	0.59	0.47
0.36	17	79	14	12	0.59	0.15	0.55	0.87	0.79	0.57	0.44
0.34	17	80	13	12	0.59	0.14	0.57	0.87	0.80	0.58	0.45
0.32	17	81	12	12	0.59	0.13	0.59	0.87	0.80	0.59	0.46
0.27	16	81	12	13	0.55	0.13	0.57	0.86	0.80	0.56	0.42
0.20	15	82	11	14	0.52	0.12	0.58	0.85	0.80	0.55	0.40
0.19	15	83	10	14	0.52	0.11	0.60	0.86	0.80	0.56	0.41
0.17	15	85	8	14	0.52	0.09	0.65	0.86	0.82	0.58	0.43
0.13	14	87	6	15	0.48	0.06	0.70	0.85	0.83	0.57	0.42
0.12	13	87	6	16	0.45	0.06	0.68	0.84	0.82	0.54	0.38
0.11	12	88	5	17	0.41	0.05	0.71	0.84	0.82	0.52	0.36
0.10	10	88	5	19	0.34	0.05	0.67	0.82	0.80	0.45	0.29
0.07	10	89	4	19	0.34	0.04	0.71	0.82	0.81	0.47	0.30
0.03	9	90	3	20	0.31	0.03	0.75	0.82	0.81	0.44	0.28
0.02	8	91	2	21	0.28	0.02	0.80	0.81	0.81	0.41	0.25
0.00	0	93	0	29	0.00	0.00	0.00	0.76	0.76	0.00	0.00

Table B.30: Window: [7.0, 14.0]. 'Acc': Accuracy. Probabilistic thresholds c_j which are repeated are ommitted and the greater presented.

Appendix C

Supplementary Figures

Here, additional figures are presented to accompany Chapters 3, 5, and 7.

C.1 Additional results from Chapter 3

C.1.1 Bias of parameter estimates under different lengths of follow-up period r

Figure C.1 provides graphical representation for the biases observed in the simulation study carried out in Section 3.4.4 already tabulated in Appendix B.1.3. We note *generally* that the parameter estimates are less biased when a longer maximal profile length, r, is available.

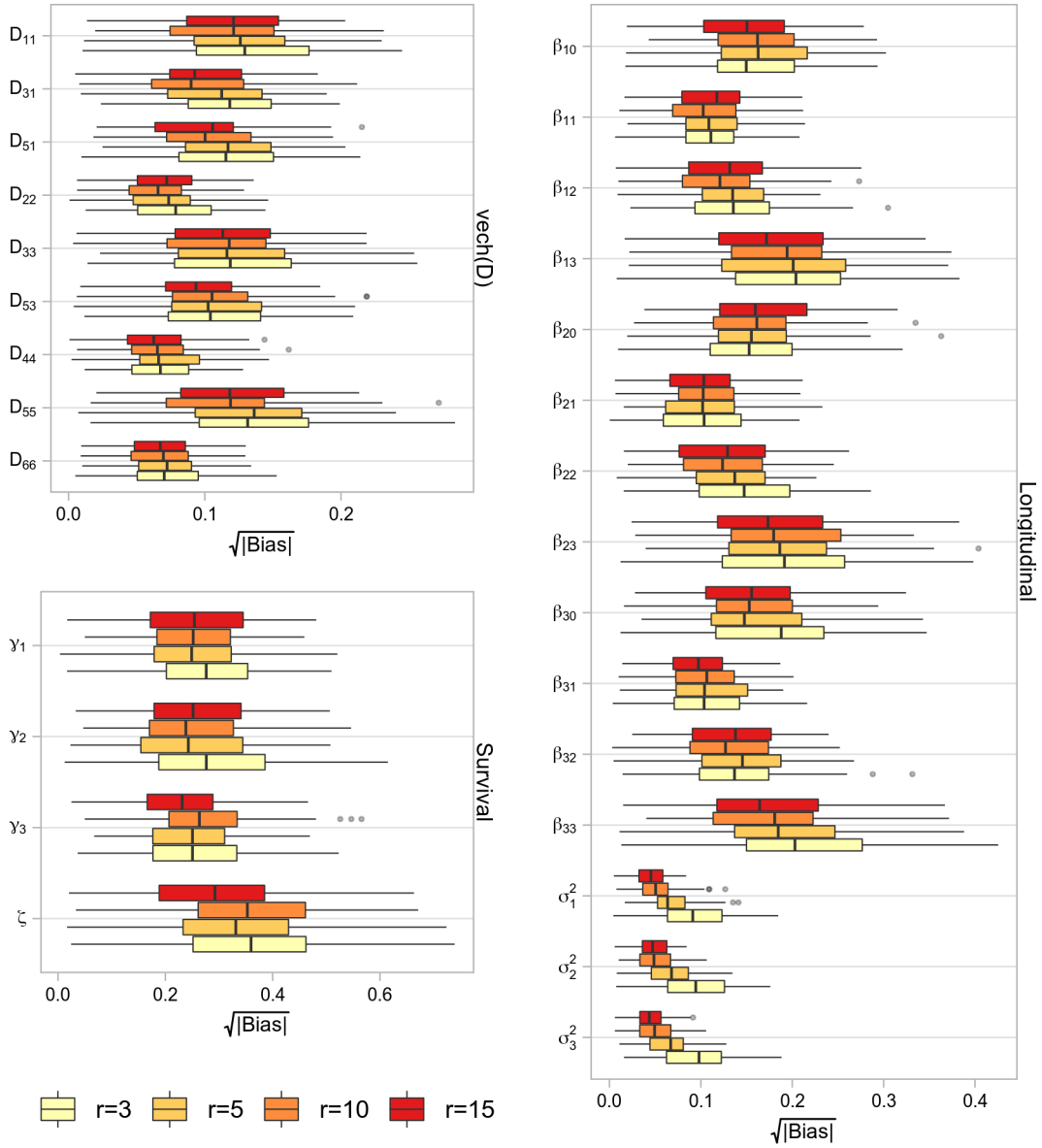


Figure C.1: Square root of absolute biases for all parameters in the simulation study into lengths of follow-up, r, carried out in Section 3.4.4. Note that these biases 'match' with those tabulated in Appendix B.1.3.

C.2 Additional results from Chapter 5

C.2.1 Scatterplot visualisations for Section 5.2.1

The families presented in the thesis are presented here in alphabetical order across Figures C.2–C.7. Each panel is created using the strategy outlined in Section 5.1.3. The random effects are sampled from the distribution $f\left(\boldsymbol{b}_i|\mathcal{D}_i;\boldsymbol{\Omega}^{(\text{TRUE})}\right)$ using a Metropolis-Hastings scheme; resultant posterior densities are discarded if the acceptance rate was not between 0.20 and 0.25. The random slopes are plotted against the random intercepts on a 'per-sample' basis. A green ellipse is generated representing where 95% of the data generated from the theoretical distribution $N\left(\hat{\boldsymbol{b}}_i,\hat{\Sigma}_i\right)$ lie using `stat_ellipse` (Wickham, 2016) and overlaid. We additionally present, as panel titles, the profile length for the subject, m_i, as well as the proportion of the scatter calculated to lie within the ellipse, ψ_i, using the method presented in Appendix A.7.

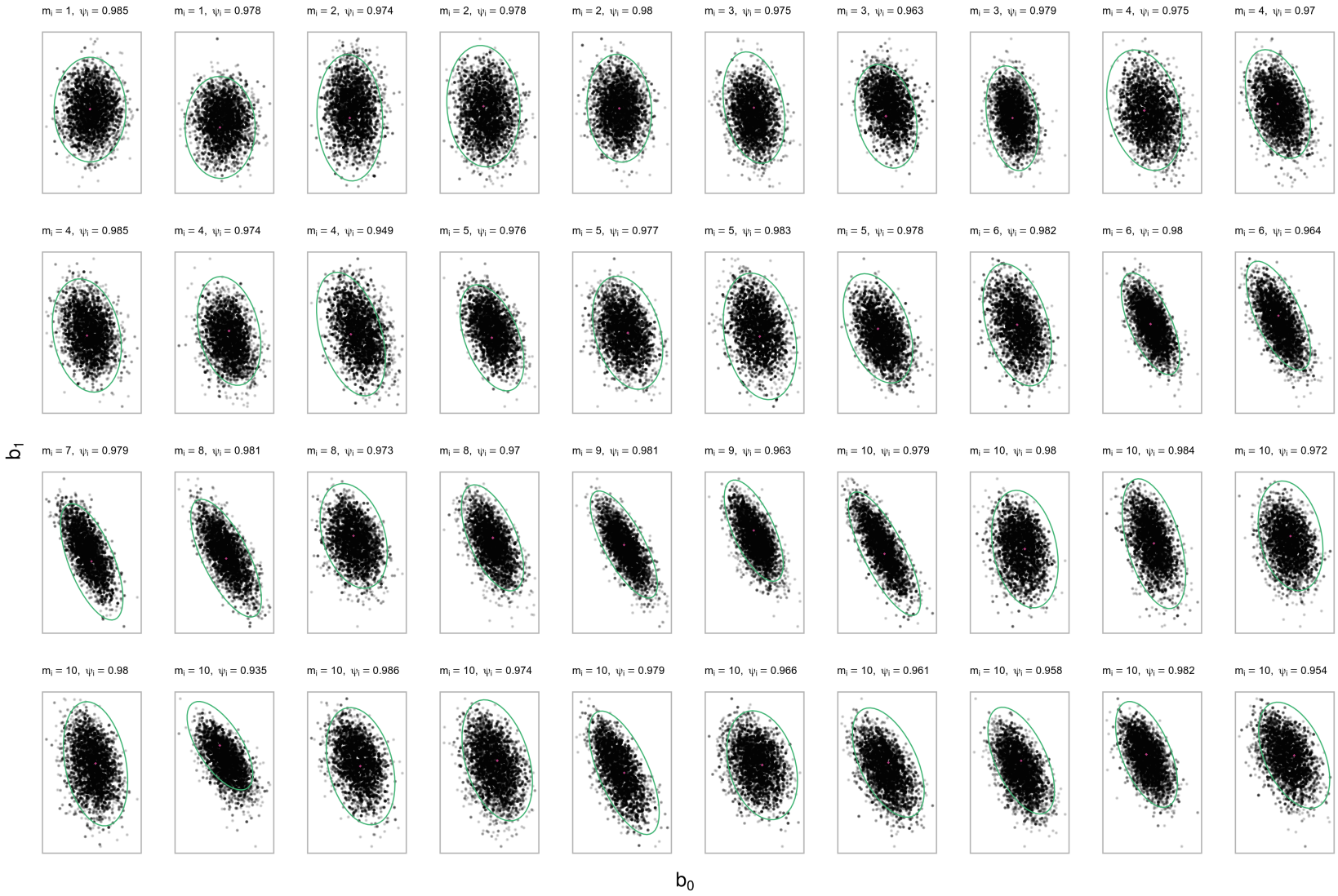


Figure C.2: Scatterplot of the sample $f\left(\boldsymbol{b}_i|\mathcal{D}_i;\boldsymbol{\Omega}^{(\text{TRUE})}\right)$ with overlaid ellipse showing the theoretical distribution $N\left(\hat{\boldsymbol{b}}_i,\hat{\Sigma}_i\right)$ for $\boldsymbol{Y}_i|\boldsymbol{b}_i;\boldsymbol{\Omega}^{(\text{TRUE})} \sim \text{Bin}(\cdot)$.

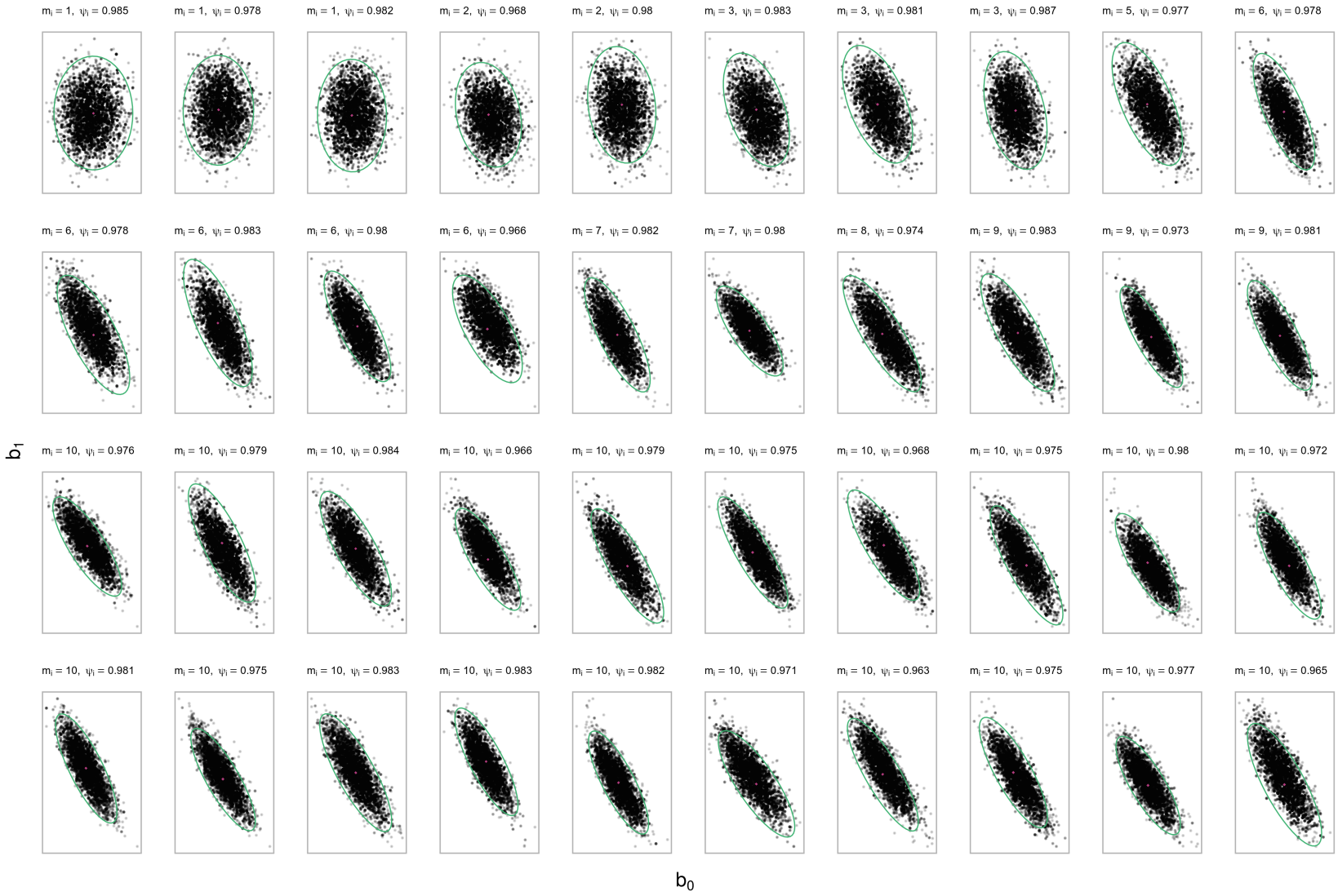


Figure C.3: Scatterplot of the sample $f\left(\boldsymbol{b}_i|\mathcal{D}_i;\boldsymbol{\Omega}^{(\text{TRUE})}\right)$ with overlaid ellipse showing the theoretical distribution $N\left(\hat{\boldsymbol{b}}_i,\hat{\Sigma}_i\right)$ for $\boldsymbol{Y}_i|\boldsymbol{b}_i;\boldsymbol{\Omega}^{(\text{TRUE})} \sim \text{Ga}(\cdot)$.

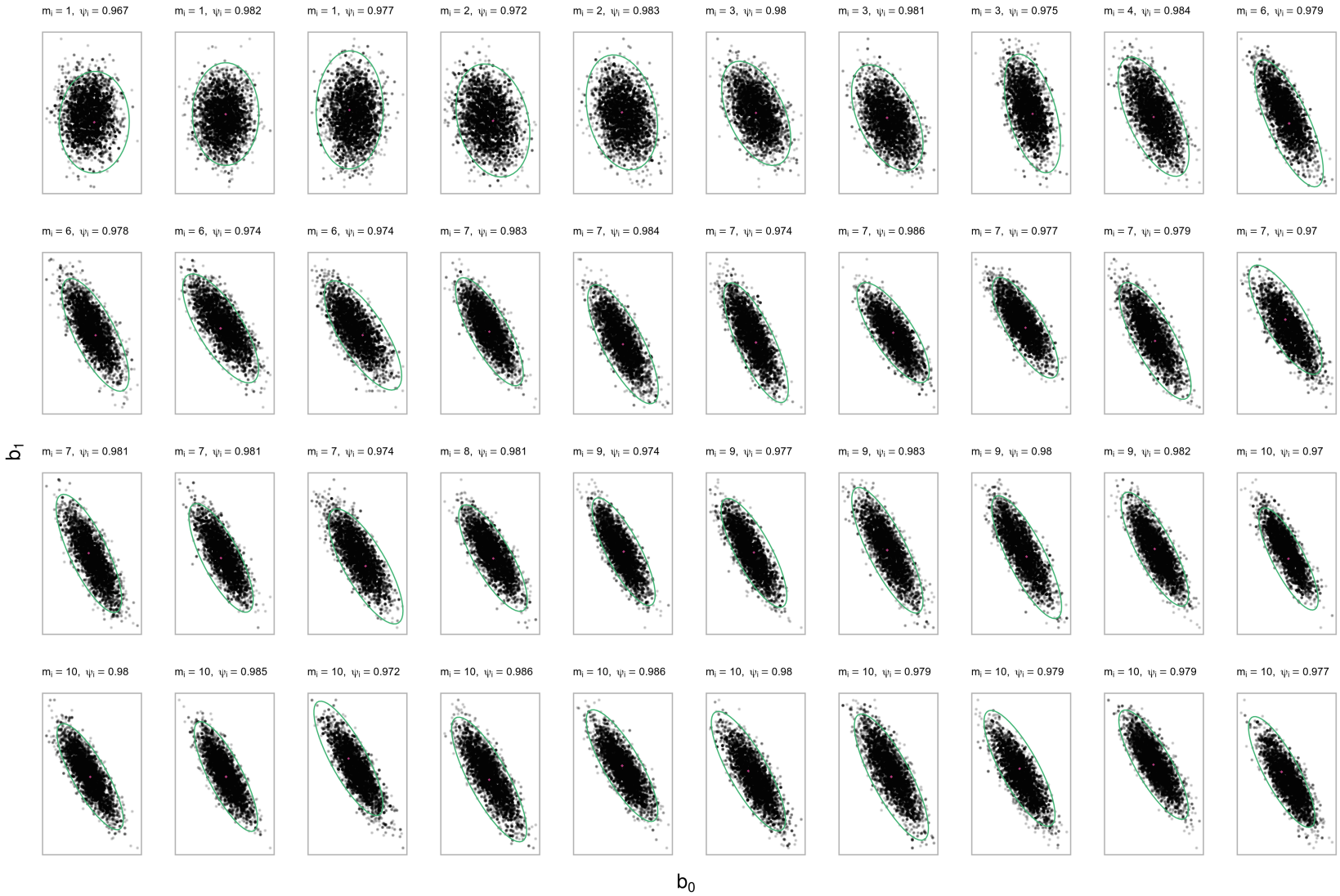


Figure C.4: Scatterplot of the sample $f\left(\boldsymbol{b}_i|\mathcal{D}_i;\boldsymbol{\Omega}^{(\text{TRUE})}\right)$ with overlaid ellipse showing the theoretical distribution $N\left(\hat{\boldsymbol{b}}_i,\hat{\Sigma}_i\right)$ for $\boldsymbol{Y}_i|\boldsymbol{b}_i;\boldsymbol{\Omega}^{(\text{TRUE})} \sim N(\cdot,\cdot)$.

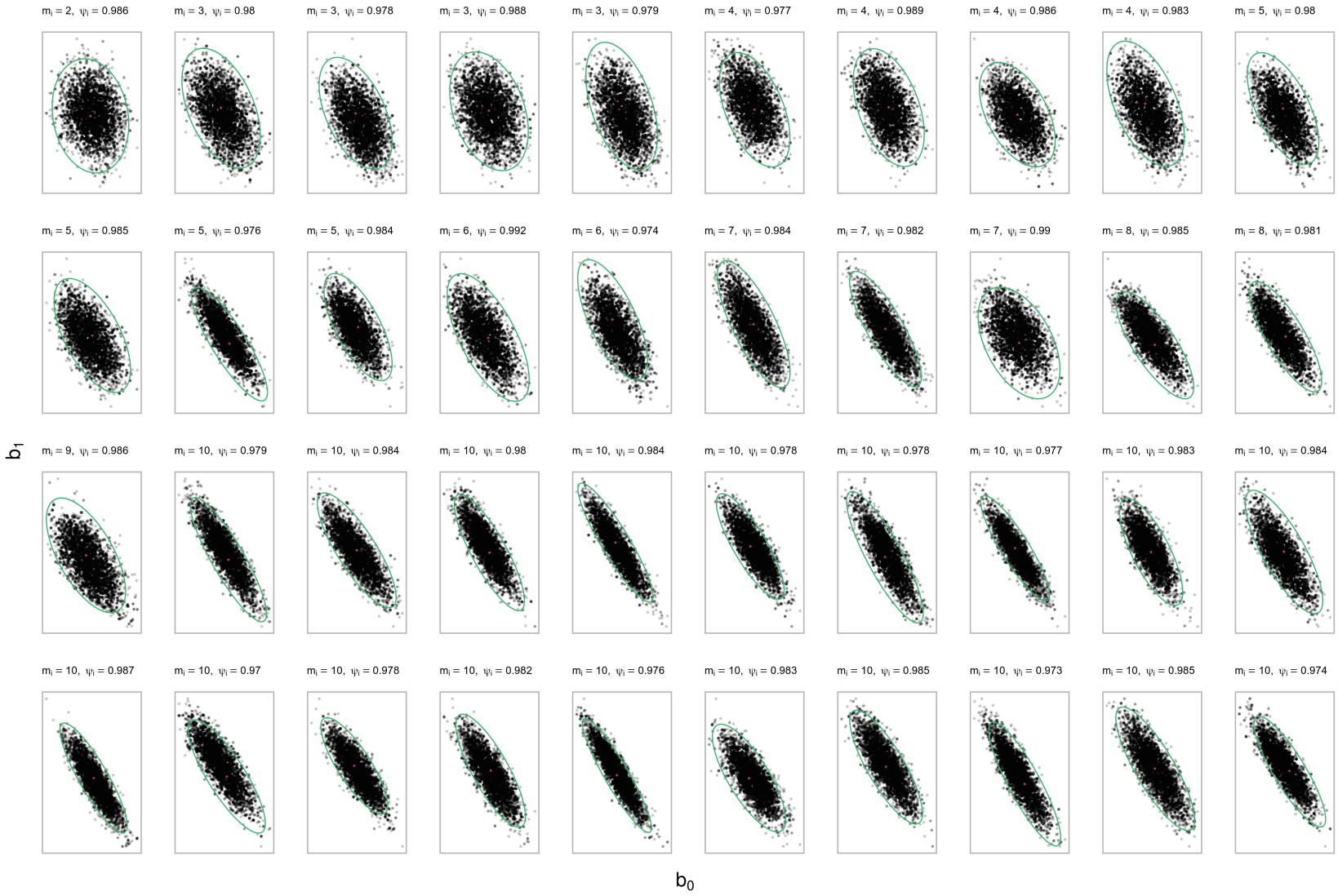


Figure C.5: Scatterplot of the sample $f\left(\boldsymbol{b}_i|\mathcal{D}_i;\boldsymbol{\Omega}^{(\text{TRUE})}\right)$ with overlaid ellipse showing the theoretical distribution $N\left(\hat{\boldsymbol{b}}_i,\hat{\Sigma}_i\right)$ for $\boldsymbol{Y}_i|\boldsymbol{b}_i;\boldsymbol{\Omega}^{(\text{TRUE})} \sim \text{GP}(\cdot)$.

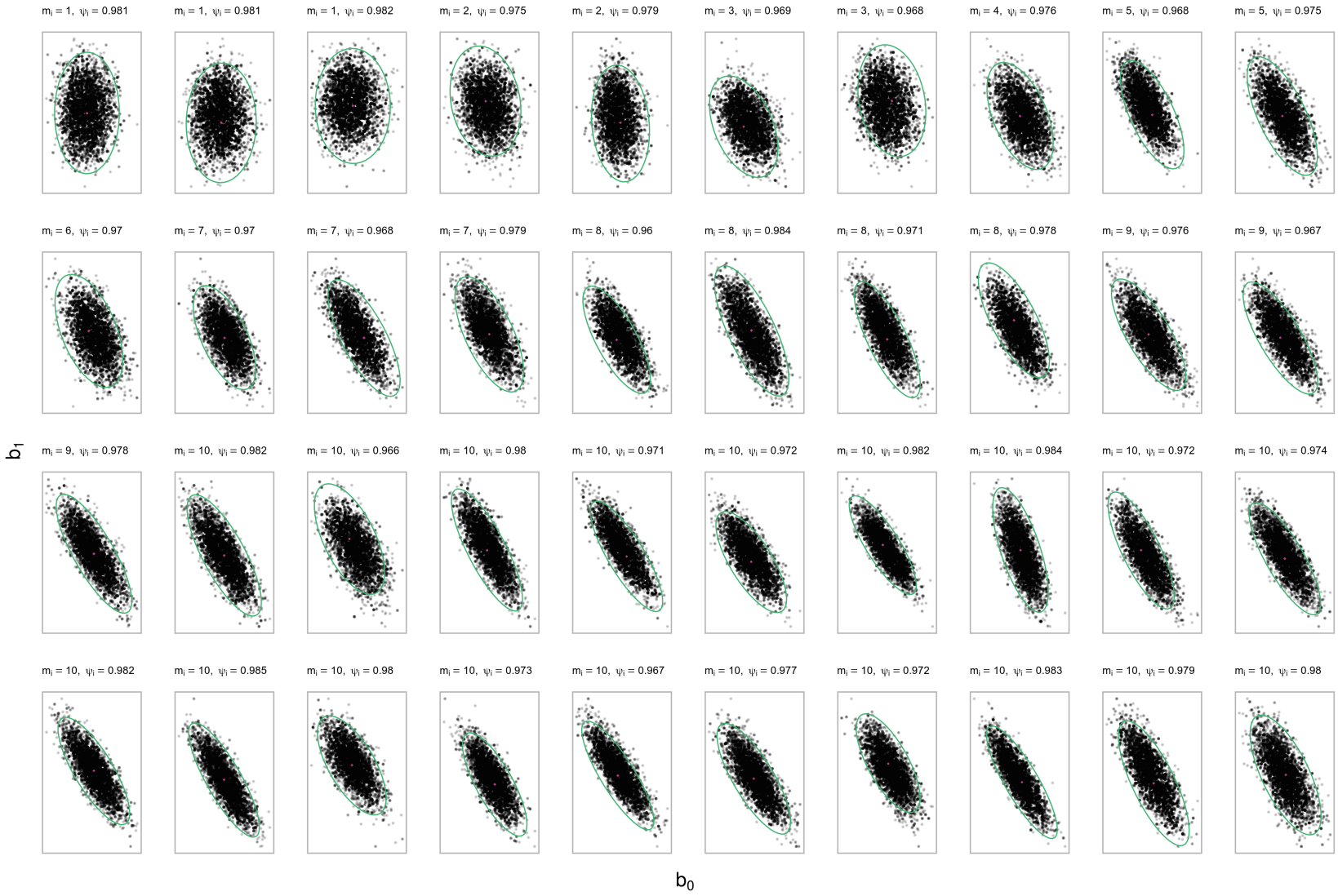


Figure C.6: Scatterplot of the sample $f\left(\boldsymbol{b}_i|\mathcal{D}_i;\boldsymbol{\Omega}^{(\text{TRUE})}\right)$ with overlaid ellipse showing the theoretical distribution $N\left(\hat{\boldsymbol{b}}_i,\hat{\Sigma}_i\right)$ for $\boldsymbol{Y}_i|\boldsymbol{b}_i;\boldsymbol{\Omega}^{(\text{TRUE})} \sim \text{NegBin}(\cdot)$.

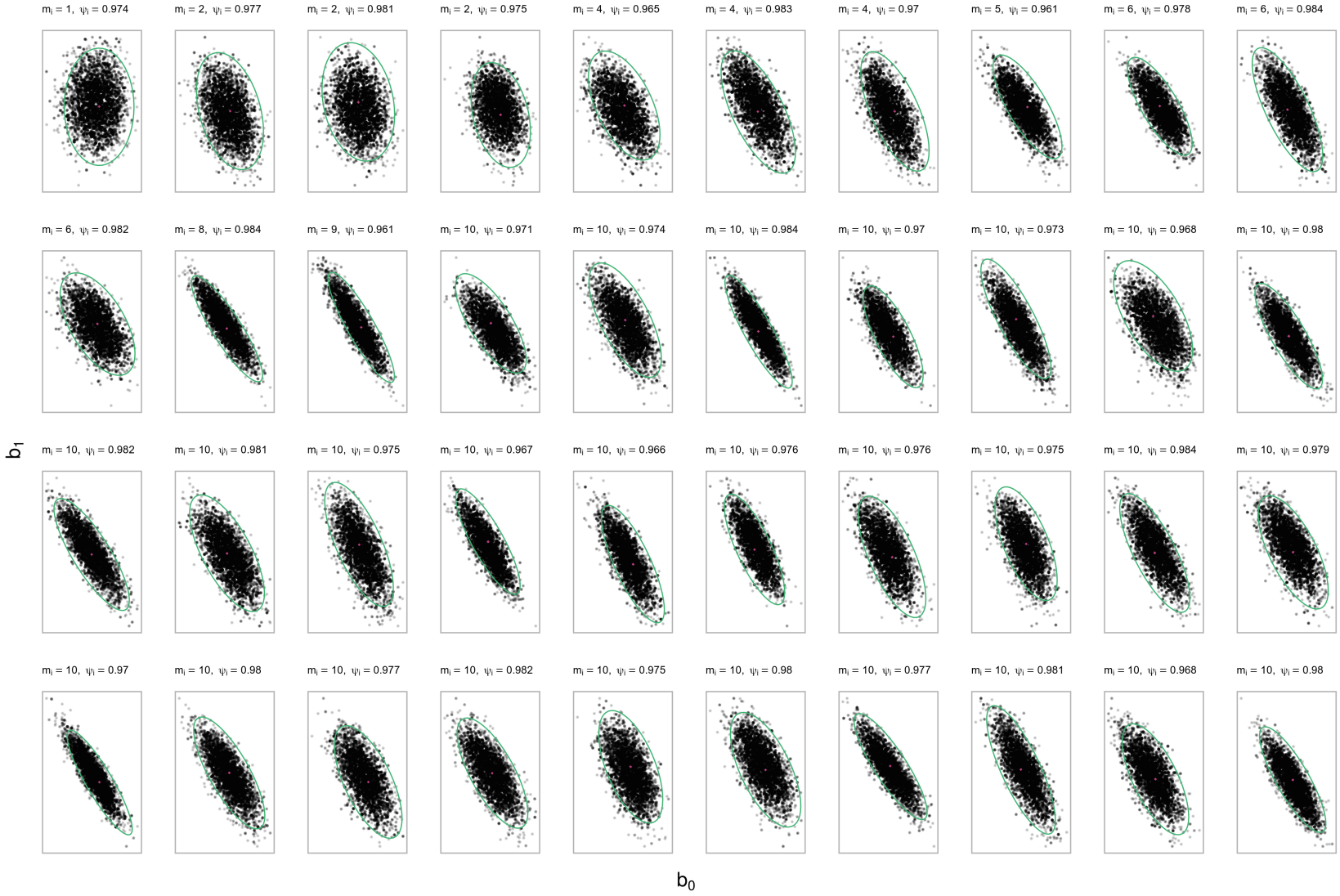


Figure C.7: Scatterplot of the sample $f\left(\boldsymbol{b}_i|\mathcal{D}_i;\boldsymbol{\Omega}^{(\text{TRUE})}\right)$ with overlaid ellipse showing the theoretical distribution $N\left(\hat{\boldsymbol{b}}_i,\hat{\Sigma}_i\right)$ for $\boldsymbol{Y}_i|\boldsymbol{b}_i;\boldsymbol{\Omega}^{(\text{TRUE})} \sim \text{Po}(\cdot)$.

C.2.2 Effect of the survival density and profile length on $\hat{\boldsymbol{b}}_i$ and the elliptical quantities $r_x,\ r_y$

In Section 5.2.3 we presented an investigation into the behaviour of the normal approximation under differing longitudinal profile lengths m_i as well as with-and-without the inclusion of the survival density in the calculation of $\hat{\boldsymbol{b}}_i$ and $\hat{\Sigma}_i$. Here we present the remaining families in alphabetical order in Figures C.8–C.11; discussion is given in Section 5.2.3.

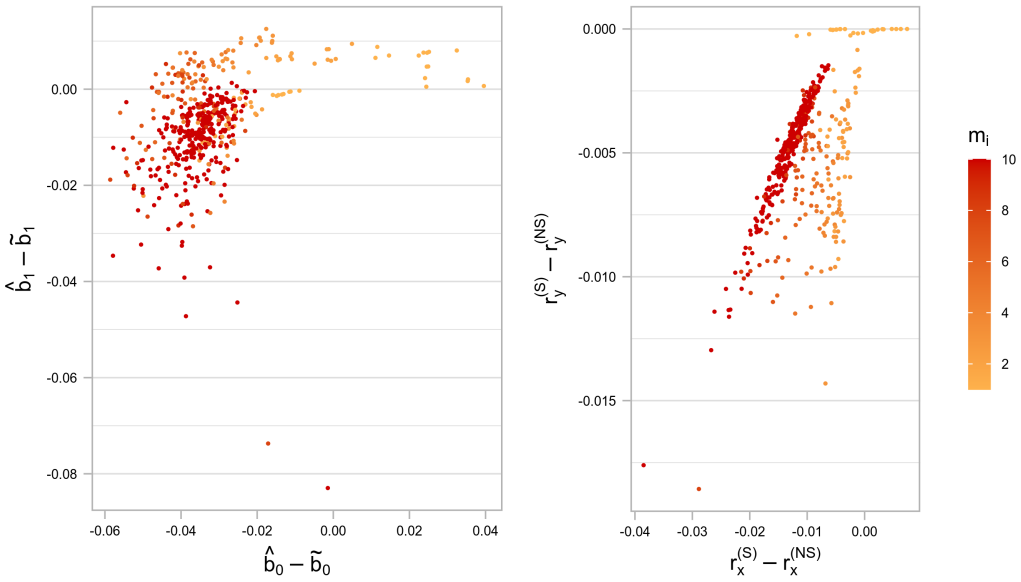


Figure C.8: Difference in the modal estimates $\hat{\boldsymbol{b}}_i - \tilde{\boldsymbol{b}}_i$; semi-minor axis $r_y^{(\mathrm{S})} - r_y^{(\mathrm{NS})}$; and semi-major $r_x^{(\mathrm{S})} - r_x^{(\mathrm{NS})}$. The differences themselves arise from the *removal* of the survival density from the complete data log-likelihood in the process to obtain the modal estimate and its covariance for $\boldsymbol{Y}_i|\boldsymbol{b}_i \sim \mathrm{Ga}(\cdot)$.

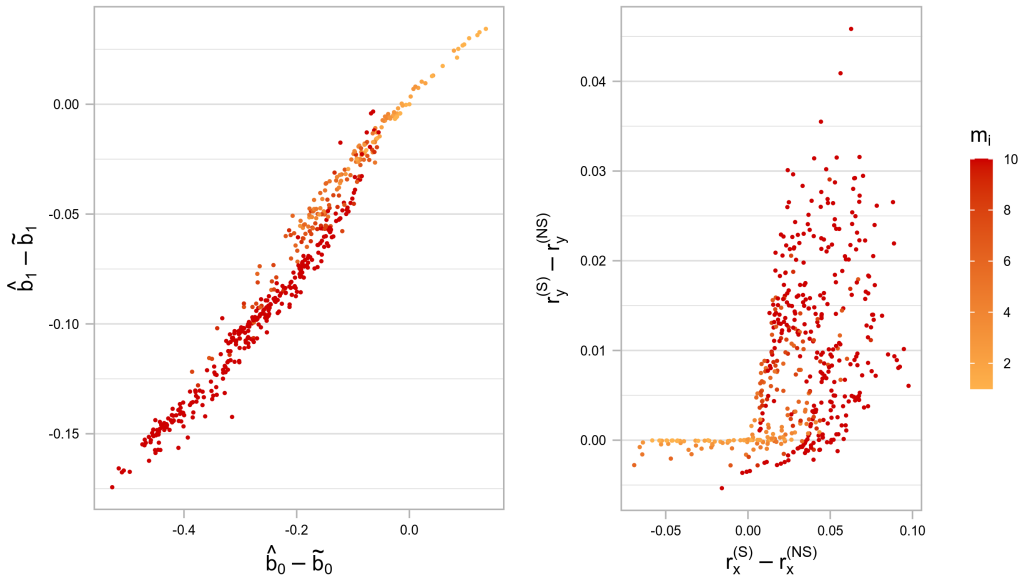


Figure C.9: Difference in the modal estimates $\hat{\boldsymbol{b}}_i - \tilde{\boldsymbol{b}}_i$; semi-minor axis $r_y^{(\mathrm{S})} - r_y^{(\mathrm{NS})}$; and semi-major $r_x^{(\mathrm{S})} - r_x^{(\mathrm{NS})}$. The differences themselves arise from the *removal* of the survival density from the complete data log-likelihood in the process to obtain the modal estimate and its covariance for $\boldsymbol{Y}_i|\boldsymbol{b}_i \sim \mathrm{GP}(\cdot)$.

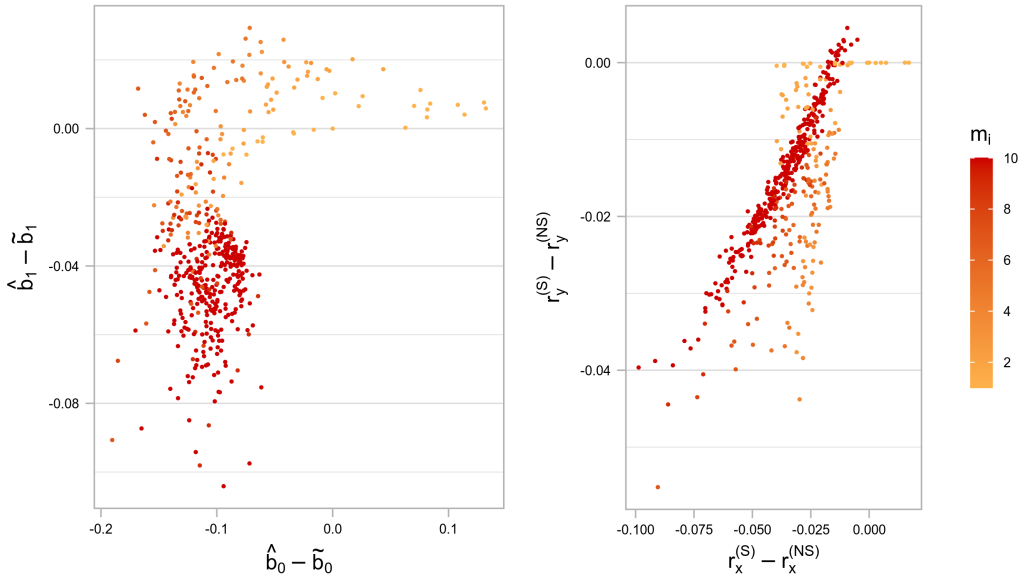


Figure C.10: Difference in the modal estimates $\hat{\boldsymbol{b}}_i - \tilde{\boldsymbol{b}}_i$; semi-minor axis $r_y^{(\mathrm{S})} - r_y^{(\mathrm{NS})}$; and semi-major $r_x^{(\mathrm{S})} - r_x^{(\mathrm{NS})}$. The differences themselves arise from the *removal* of the survival density from the complete data log-likelihood in the process to obtain the modal estimate and its covariance for $\boldsymbol{Y}_i|\boldsymbol{b}_i \sim \mathrm{NegBin}(\cdot)$.

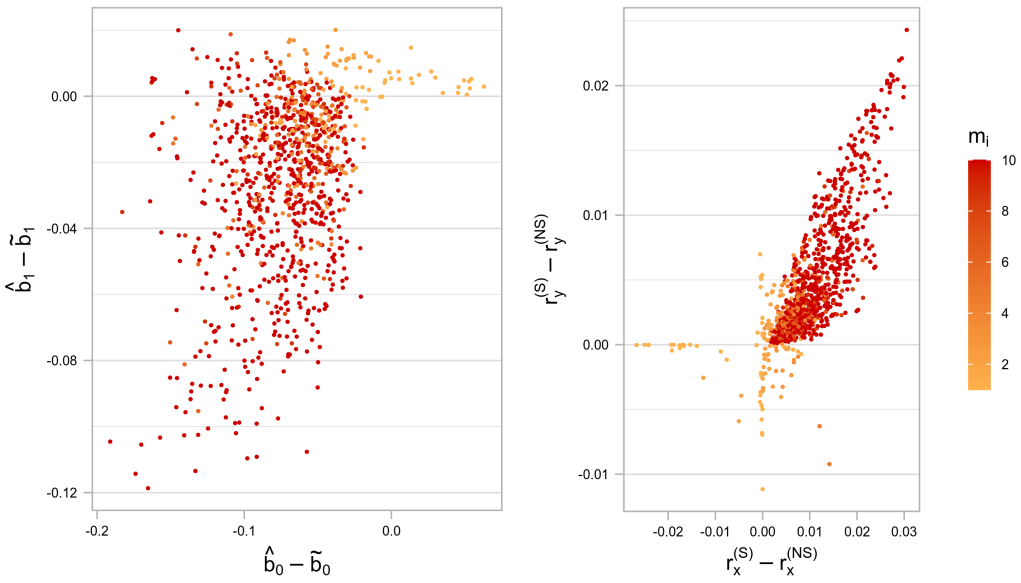


Figure C.11: Difference in the modal estimates $\hat{\boldsymbol{b}}_i - \tilde{\boldsymbol{b}}_i$; semi-minor axis $r_y^{(\mathrm{S})} - r_y^{(\mathrm{NS})}$; and semi-major $r_x^{(\mathrm{S})} - r_x^{(\mathrm{NS})}$. The differences themselves arise from the *removal* of the survival density from the complete data log-likelihood in the process to obtain the modal estimate and its covariance for $\boldsymbol{Y}_i | \boldsymbol{b}_i \sim \mathrm{Po}(\cdot)$.

C.2.3 Posterior density of $g(\boldsymbol{b}_i)\,|T_i, \Delta_i, \boldsymbol{Y}_i; \hat{\boldsymbol{\Omega}} = \log(\mathbf{1} + \exp\{\mathrm{X}_i\boldsymbol{\beta} + \mathrm{Z}_i\boldsymbol{b}_i\})$

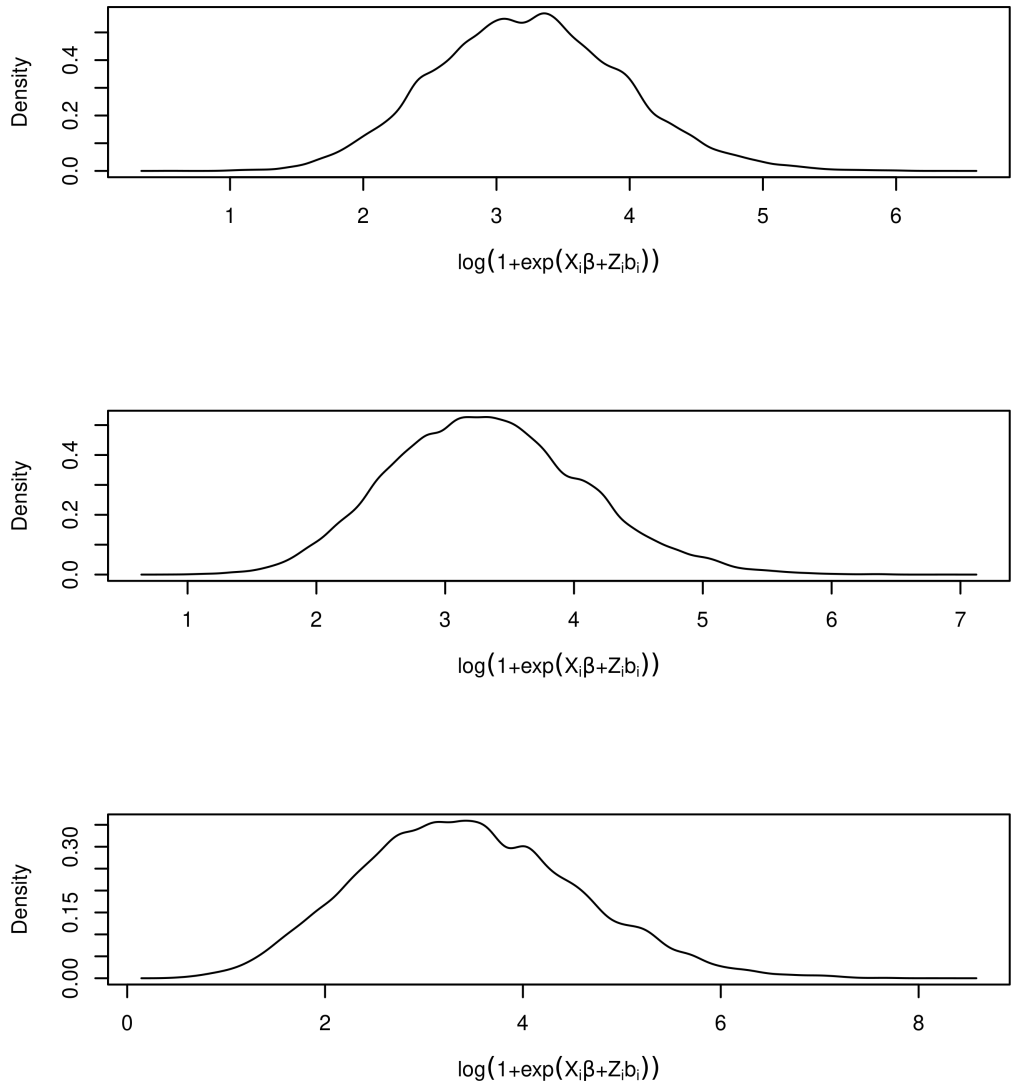


Figure C.12: Posterior density for $g(\boldsymbol{b}_i)\,|T_i, \Delta_i, \boldsymbol{Y}_i; \hat{\boldsymbol{\Omega}} = \log(\mathbf{1} + \exp\{\mathrm{X}_i\boldsymbol{\beta} + \mathrm{Z}_i\boldsymbol{b}_i\})$ evaluated at the first (top pane), middle and final (bottom pane) follow-up time for univariate binomial simulated data.

C.2.4 Posterior density of $g(\boldsymbol{b}_i)\,|T_i, \Delta_i, \boldsymbol{Y}_i; \hat{\boldsymbol{\Omega}} = \log(\boldsymbol{Y}_i\boldsymbol{\varphi}_i + \exp\{\mathrm{X}_i\boldsymbol{\beta} + \mathrm{Z}_i\boldsymbol{b}_i\})$

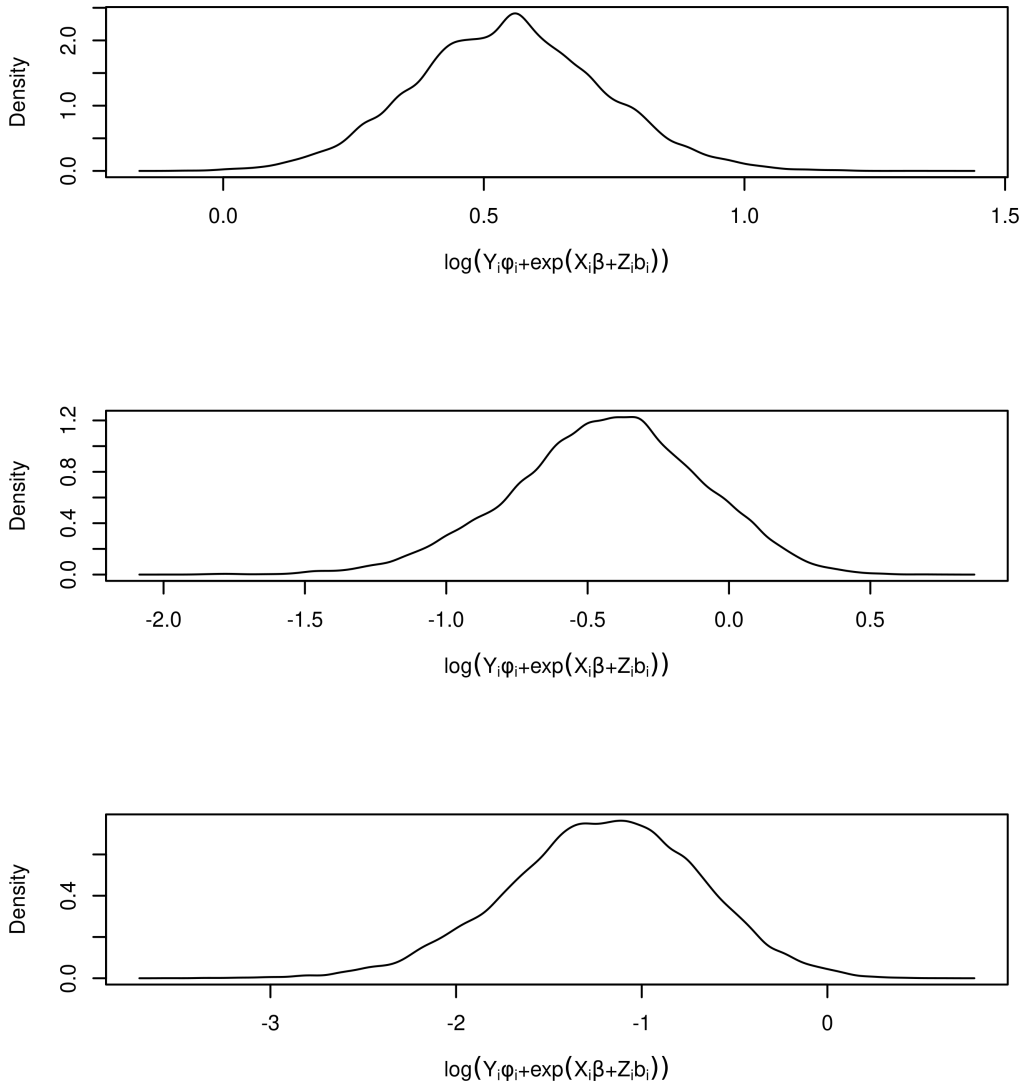


Figure C.13: Posterior density for $g(\boldsymbol{b}_i)\,|T_i, \Delta_i, \boldsymbol{Y}_i; \hat{\boldsymbol{\Omega}} = \log(\boldsymbol{Y}_i\boldsymbol{\varphi}_i + \exp\{\mathrm{X}_i\boldsymbol{\beta} + \mathrm{Z}_i\boldsymbol{b}_i\})$ evaluated at the first (top pane), middle and final (bottom pane) follow-up time for univariate generalised Poisson simulated data.

C.2.5 Posterior density of $g(\boldsymbol{b}_i)\,|T_i, \Delta_i, \boldsymbol{Y}_i; \hat{\boldsymbol{\Omega}} = \exp\{\mathrm{X}_i\boldsymbol{\beta} + \mathrm{Z}_i\boldsymbol{b}_i\}$ for other families

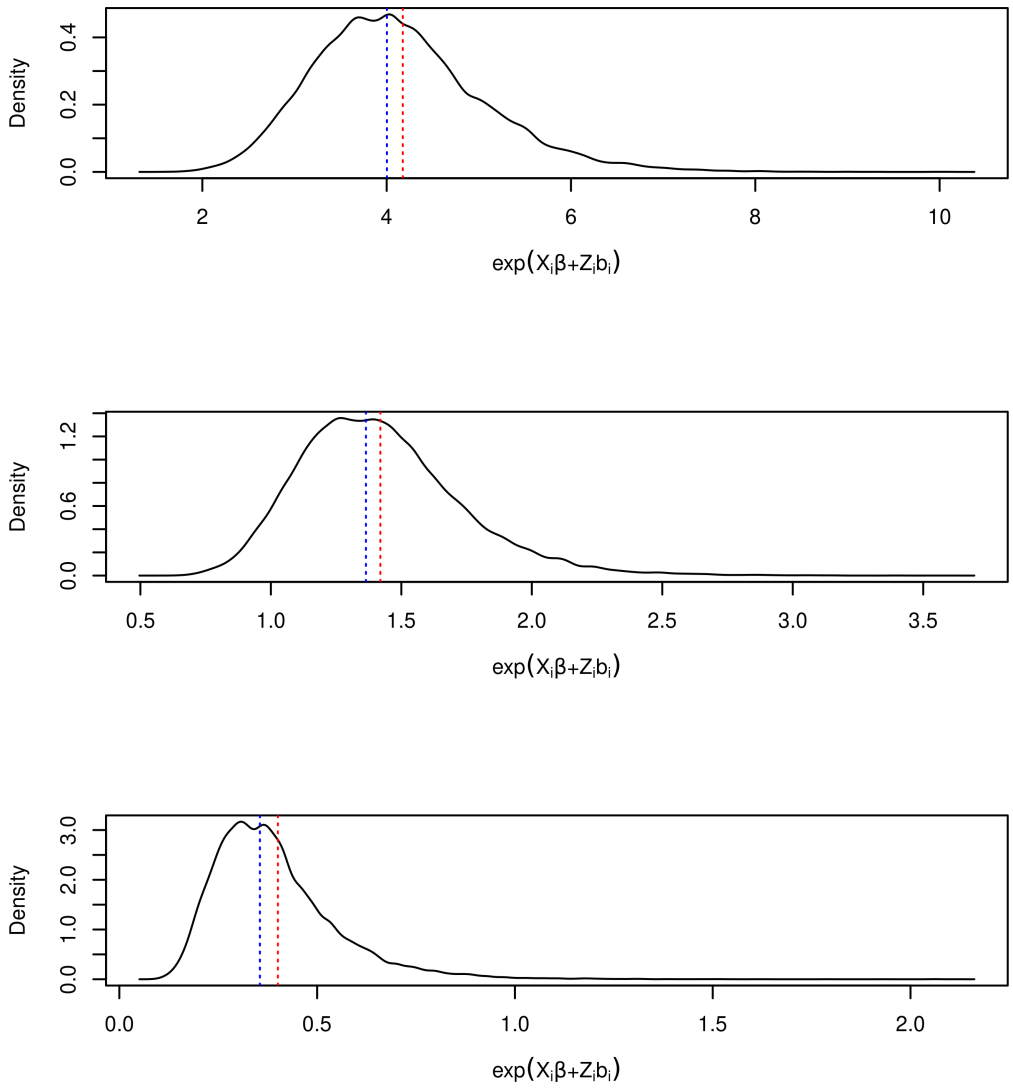


Figure C.14: Posterior density for $g(\boldsymbol{b}_i)\,|T_i, \Delta_i, \boldsymbol{Y}_i; \hat{\boldsymbol{\Omega}} = \exp\{\mathrm{X}_i\boldsymbol{\beta} + \mathrm{Z}_i\boldsymbol{b}_i\}$ evaluated at the first (top pane), middle and final (bottom pane) follow-up time for univariate Gamma simulated data. The mean and median of the posterior distribution are denoted by the red and blue dashed lines, respectively.

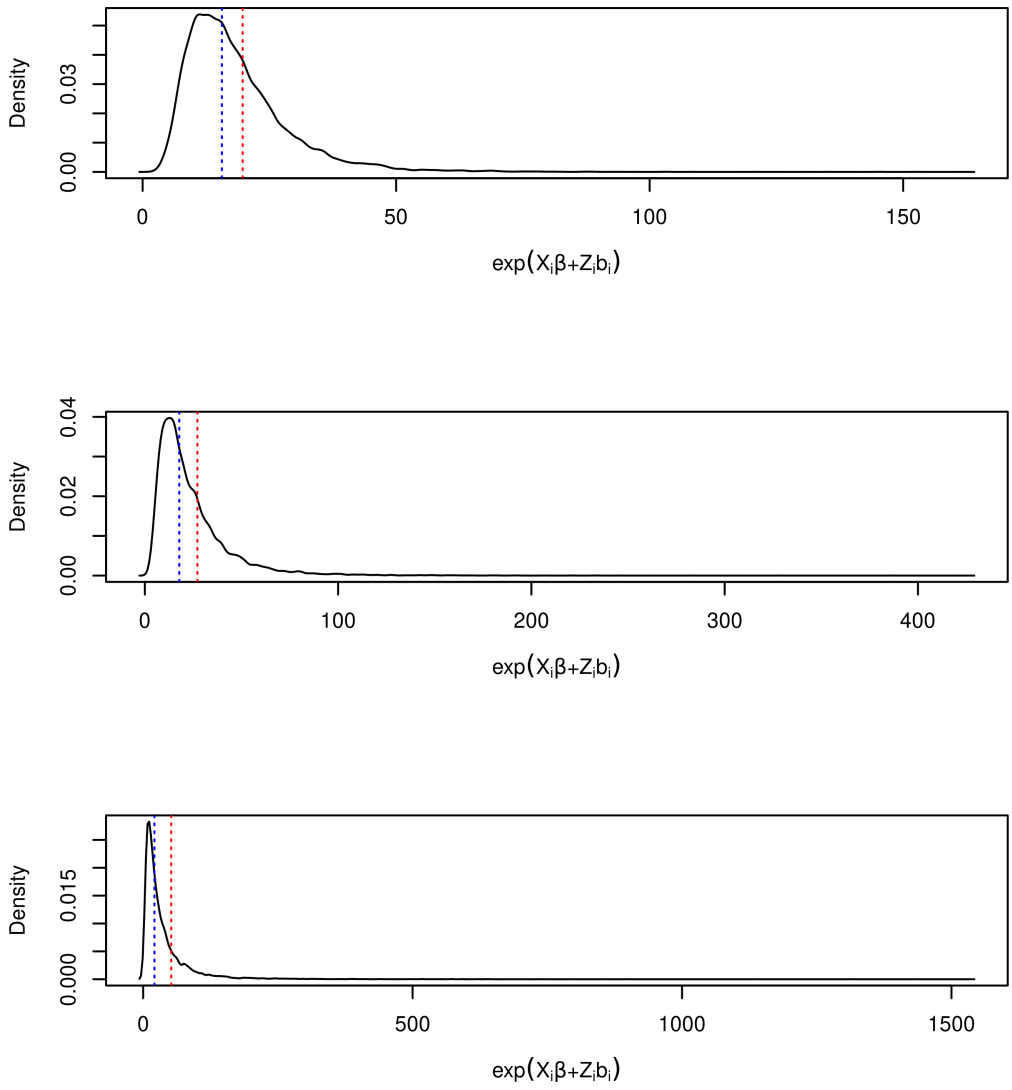


Figure C.15: Posterior density for $g(\boldsymbol{b}_i)\,|T_i, \Delta_i, \boldsymbol{Y}_i; \hat{\boldsymbol{\Omega}} = \exp\{\mathrm{X}_i\boldsymbol{\beta} + \mathrm{Z}_i\boldsymbol{b}_i\}$ evaluated at the first (top pane), middle and final (bottom pane) follow-up time for univariate binomial simulated data. The mean and median of the posterior distribution are denoted by the red and blue dashed lines, respectively.

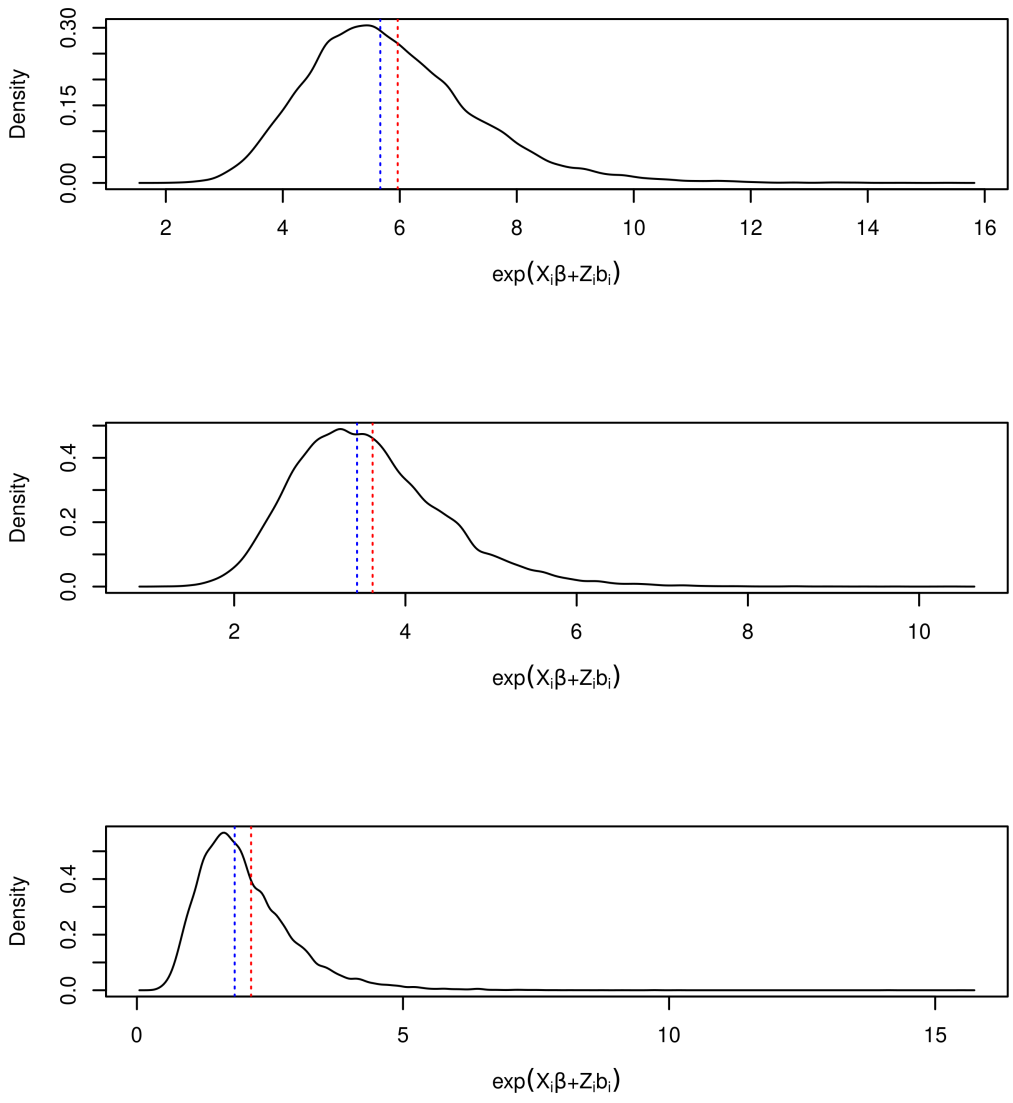


Figure C.16: Posterior density for $g(\boldsymbol{b}_i)\,|T_i, \Delta_i, \boldsymbol{Y}_i; \hat{\boldsymbol{\Omega}} = \exp\{\mathrm{X}_i\boldsymbol{\beta} + \mathrm{Z}_i\boldsymbol{b}_i\}$ evaluated at the first (top pane), middle and final (bottom pane) follow-up time for univariate negative binomial simulated data. The mean and median of the posterior distribution are denoted by the red and blue dashed lines, respectively.

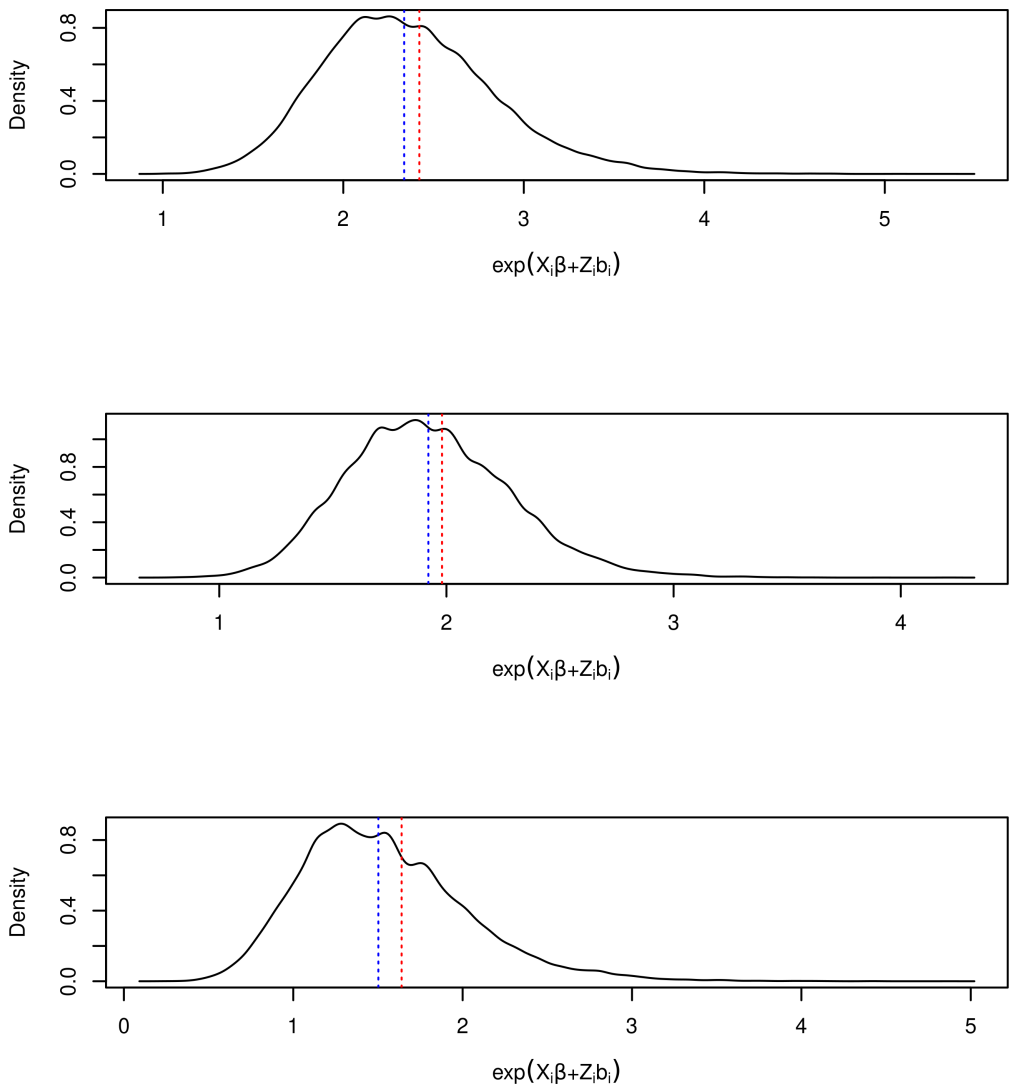


Figure C.17: Posterior density for $g(\boldsymbol{b}_i) = \exp\{\mathrm{X}_i\boldsymbol{\beta} + \mathrm{Z}_i\boldsymbol{b}_i\}\,|T_i, \Delta_i, \boldsymbol{Y}_i; \hat{\boldsymbol{\Omega}}$ evaluated at the first (top pane), middle and final (bottom pane) follow-up time for univariate generalised Poisson simulated data. The mean and median of the posterior distribution are denoted by the red and blue dashed lines, respectively.

C.2.6 Posterior density of $g(\boldsymbol{b}_i)\,|T_i, \Delta_i, \boldsymbol{Y}_i; \hat{\boldsymbol{\Omega}} = \exp\{\mathrm{X}_i\boldsymbol{\beta} + \mathrm{Z}_i\boldsymbol{b}_i\}$ for the Poisson case, with modal value

The mode of the log-normal quantity $\exp\left\{\mathrm{X}_i\boldsymbol{\beta} + \mathrm{Z}_i\hat{\boldsymbol{b}}_i\right\} \overset{\text{appx.}}{\sim} LN(\hat{\boldsymbol{\mu}}_i, \mathrm{A}_i)$ is given by $\exp\{\hat{\boldsymbol{\mu}}_i - \boldsymbol{\tau}_i^2\}$. We present this quantity in contention with the median and mean for the Poisson case only in Figure C.18.

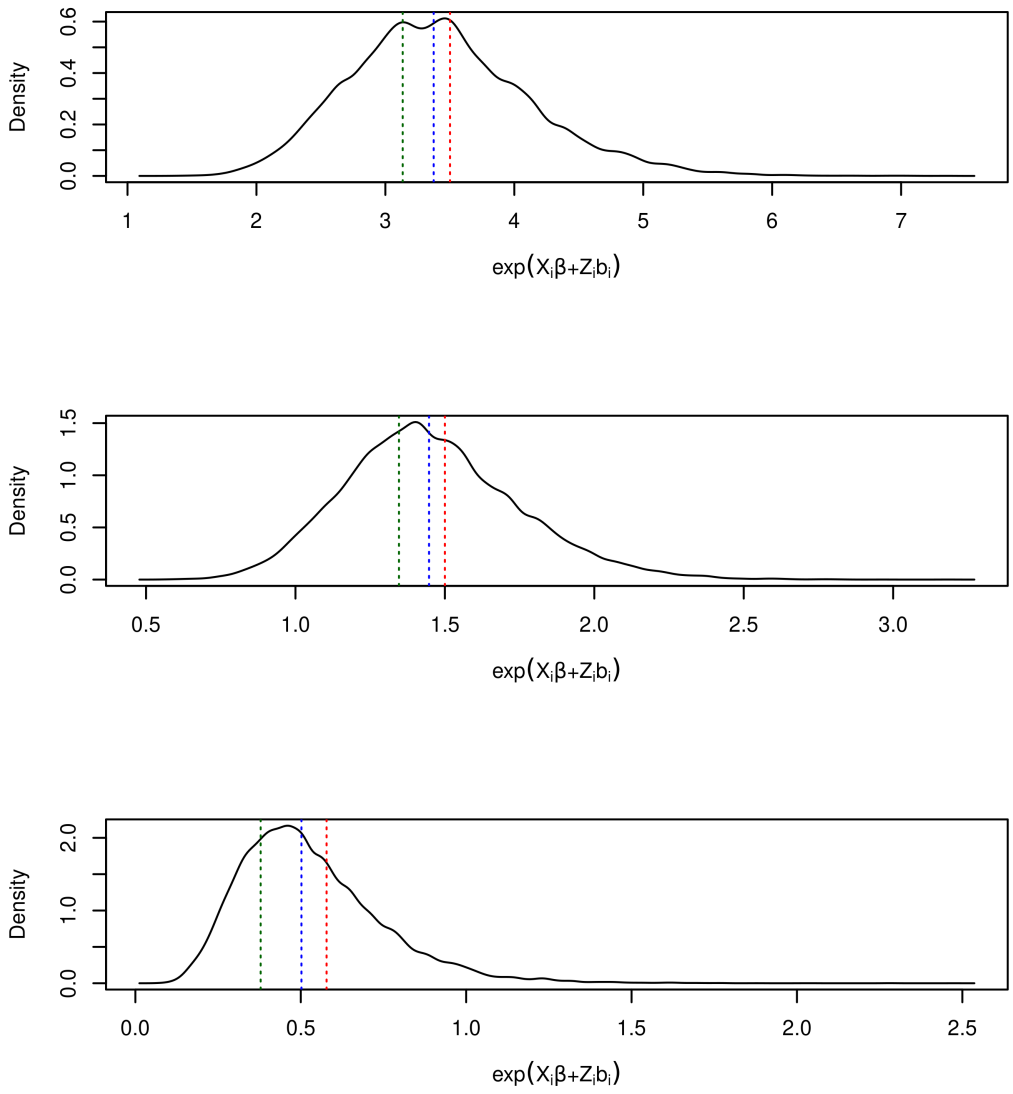


Figure C.18: Posterior density for $g(\boldsymbol{b}_i)\,|T_i, \Delta_i, \boldsymbol{Y}_i; \hat{\boldsymbol{\Omega}} = \exp\{\mathrm{X}_i\boldsymbol{\beta} + \mathrm{Z}_i\boldsymbol{b}_i\}$ evaluated at the first (top pane), middle and final (bottom pane) follow-up time for univariate Poisson simulated data. The mean, median, and mode of the posterior distribution are denoted by the red, blue, and dark green dashed lines, respectively.

C.2.7 Values for a which achieve nominal coverage in $\psi_i(\hat{\boldsymbol{b}}_i, a\hat{\Sigma}_i)$

In Section 5.5 we obtain values for a which minimised the objective function (5.6). Summaries of the one hundred values obtained for a for each of the families we consider across Chapters 2–4 was presented in Table 5.1; here we provide supplementary figures displaying *all* estimates a in alphabetical order by family.

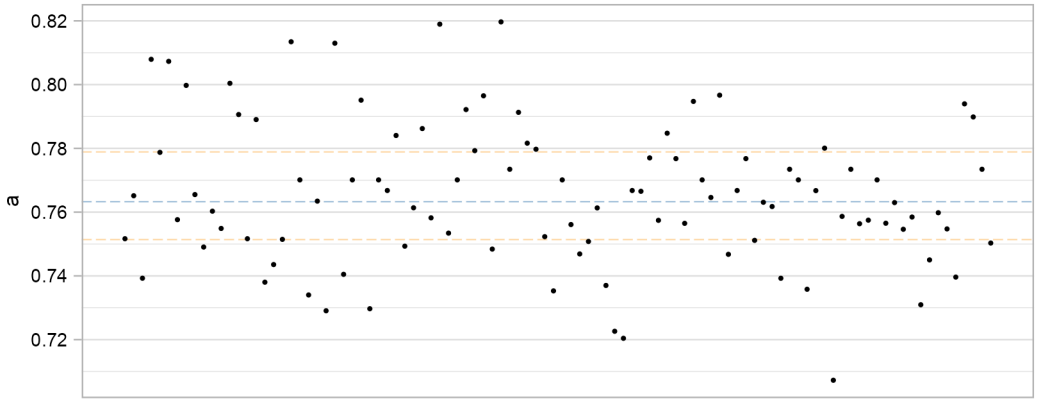


Figure C.19: Values for a which minimise $Q(a)$ in (5.6); each point represents the value a from *one* set of Binomial simulated data. The median value is denoted by the blue dashed line and the orange dashed lines represent the interquartile range.

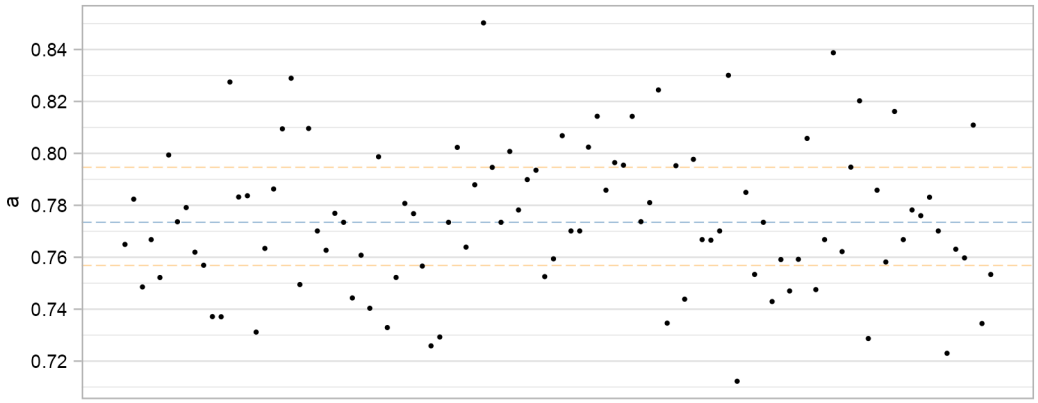


Figure C.20: Values for a which minimise $Q(a)$ in (5.6); each point represents the value a from *one* set of Gamma simulated data. The median value is denoted by the blue dashed line and the orange dashed lines represent the interquartile range.

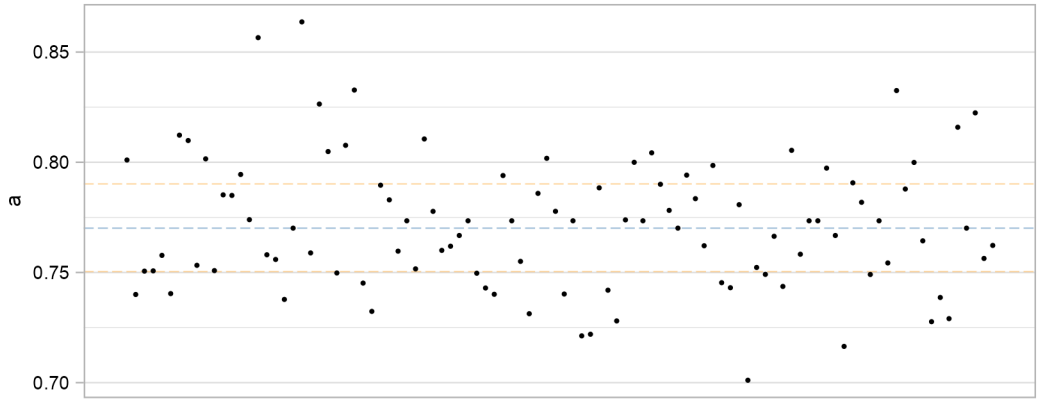


Figure C.21: Values for a which minimise $Q(a)$ in (5.6); each point represents the value a from *one* set of generalised Poisson simulated data. The median value is denoted by the blue dashed line and the orange dashed lines represent the interquartile range.

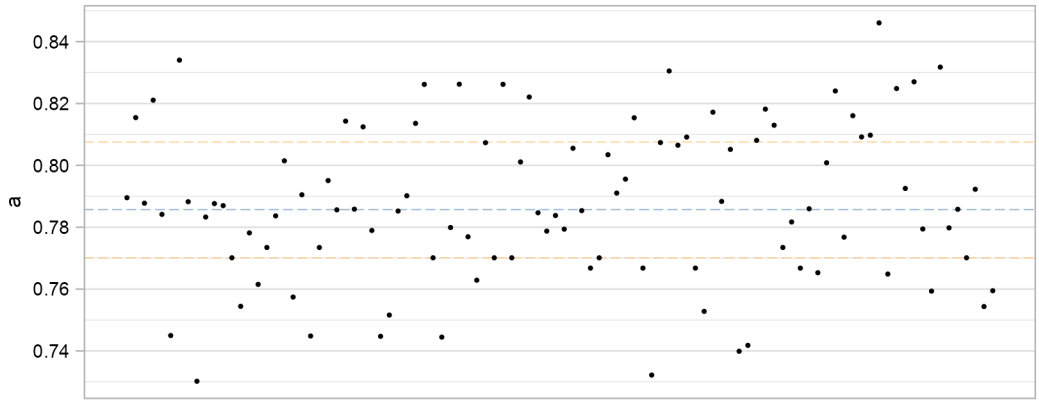


Figure C.22: Values for a which minimise $Q(a)$ in (5.6); each point represents the value a from *one* set of negative binomial simulated data. The median value is denoted by the blue dashed line and the orange dashed lines represent the interquartile range.

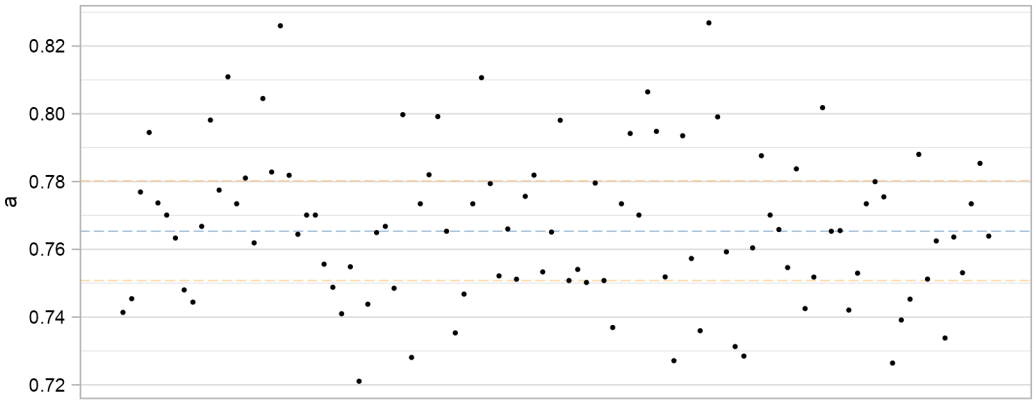


Figure C.23: Values for a which minimise $Q(a)$ in (5.6); each point represents the value a from *one* set of Poisson simulated data. The median value is denoted by the blue dashed line and the orange dashed lines represent the interquartile range.

C.3 Additional results from Chapter 7

C.3.1 Prevalence of longitudinal biomarkers over time

In Section 7.2 we presented the trajectories of continuous and count longitudinal biomarkers over follow-up. Here, we attempt to present the prevalence of the binary biomarkers (i.e. $y_{ijk} = 1$) for some time j within a percentile of follow-up. This pseudo-heatmap is presented in Figure C.24, where we can fairly easily observe the relative absence of both spiders and ascites in comparison with hepatomegaly.

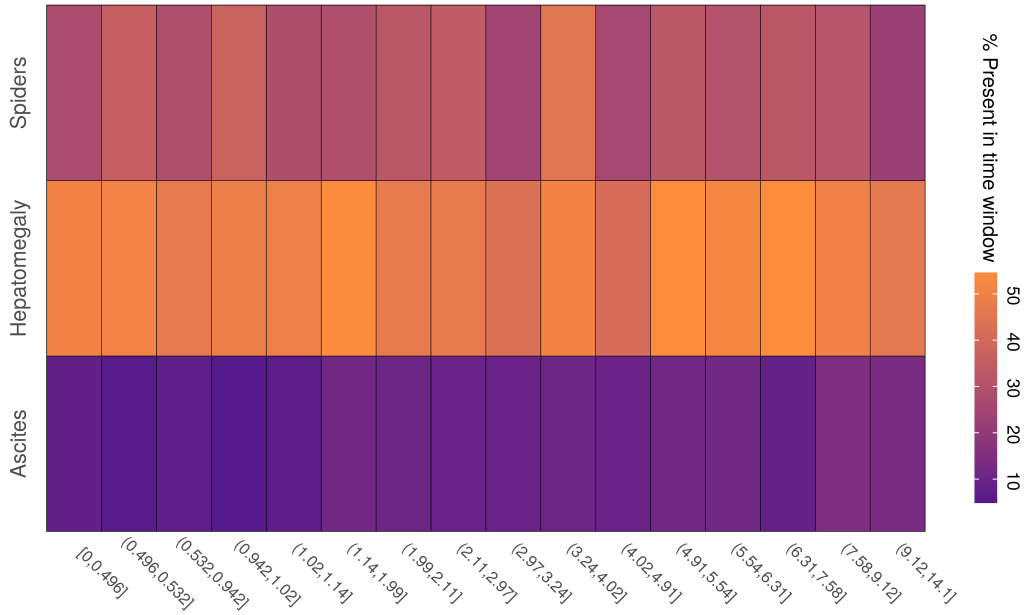


Figure C.24: Heatmap presenting the prevalence (defined as the proportion of present cases to observed cases) in each of the time-windows presented along the x-axis.

C.3.2 Dispersion model selection

In Section 7.3.2 we arrived at the best-fitting longitudinal sub-model for each response we consider, here we present further consideration of all possible (univariate) dispersion models.

C.3.3 Pearson residuals for univariate joint models

Each univariate joint model fit in Section 7.4.1 produces fitted values and corresponding Pearson residuals which we outlined in Section 6.1.1. We present these in Figure C.26

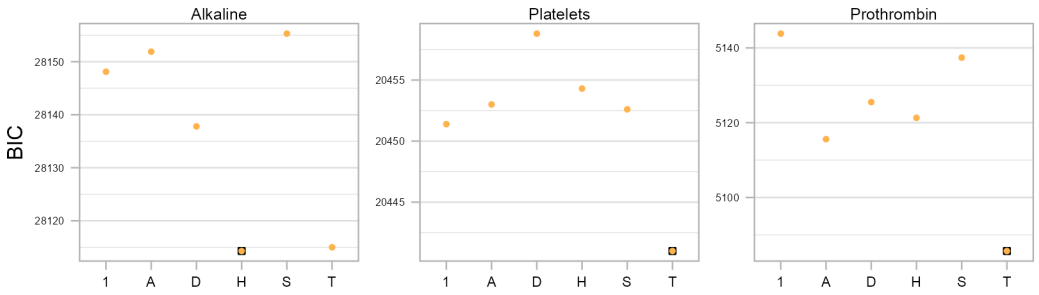


Figure C.25: BIC rankings for dispersion models for each longitudinal response. The x-axis presents the baseline covariate used in the dispersion model `A`: `age`; `D`: `drug` `H`: `histologic`; `S`: `sex`; and additionally `T`: `time`. Where the x-axis reads `1` refers to the 'intercept-only' dispersion model, which is fit by default i.e. the BIC for the model found in Section 7.3.2. A black square is drawn behind the model with the lowest BIC for each response for clarity's sake.

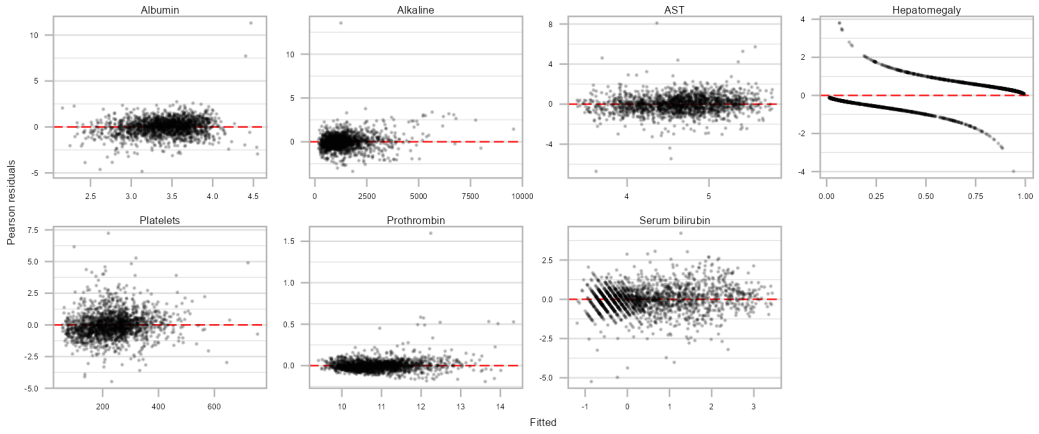


Figure C.26: Pearson residuals plotted against fitted values from each univariate joint model fit to PBC data; the biomarker is shown in the panel title.

C.3.4 Cox-Snell residuals for univariate joint models

Each univariate joint model fit in Section 7.4.1 produces Cox-Snell residuals outlined in Section 6.1.2. We note broadly good agreement between the estimate of the survival function and the unit exponential we expect to see across *all* biomarkers. We two examples only (due to there being little to distinguish visually): One where there is very good agreement (serum bilirubin) and another which exhibits far less overlap of the two curves (albumin); presented in Figure C.27.

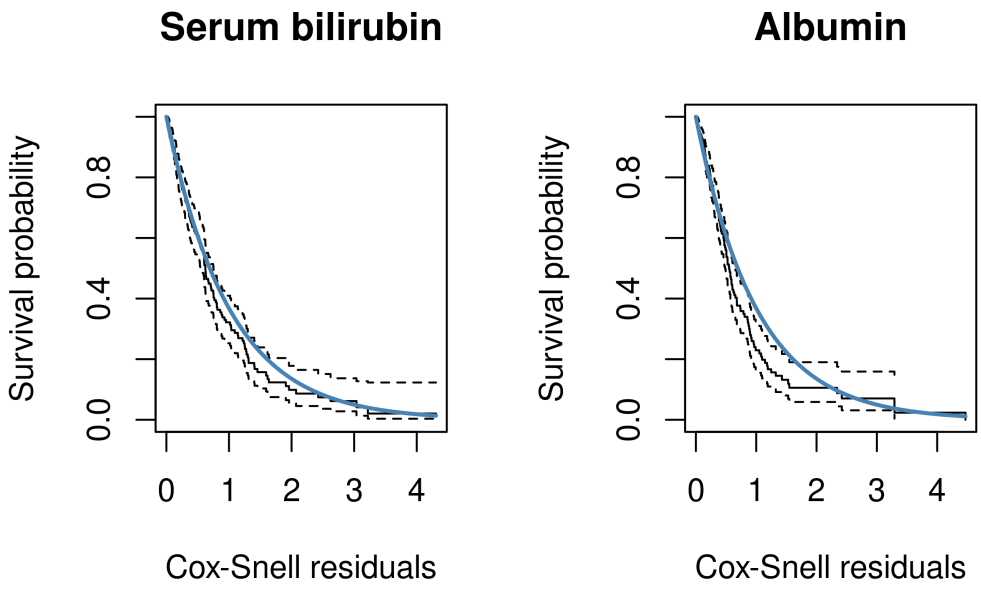


Figure C.27: Survival function of the Cox-Snell residuals obtained from the series of univariate model fits. Only two biomarkers (i.e. the residuals from two separate joint models) are shown as an example. The steel-blue overlaid curve represents the theoretical unit exponential distribution.

C.3.5 Pearson residuals from final model

Pearson residuals $\hat{r}_{ikj}^{(P)}$ against fitted values $\hat{Y}_{ikj}$ for each of the $k = 1, \ldots, K$ longitudinal sub-models are presented in Figure C.28 for the final model in Section 7.5.

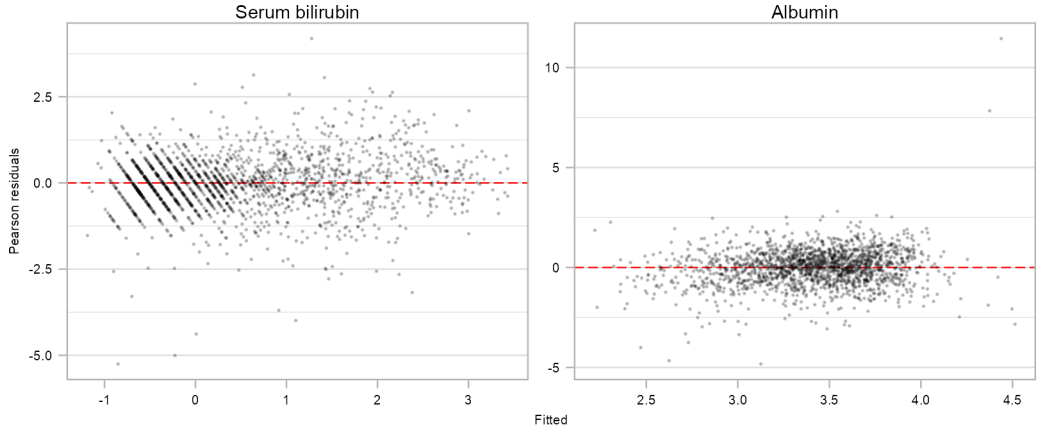


Figure C.28: Pearson residuals plotted against fitted values obtained for each longitudinal sub-model in the 'final' joint model fit to the PBC data in Section 7.5; the biomarker is shown in the panel title.

C.3.6 Cox-Snell residuals from final model

Cox-Snell residuals for the multivariate joint model in Section 7.5 are presented in Figure C.29.

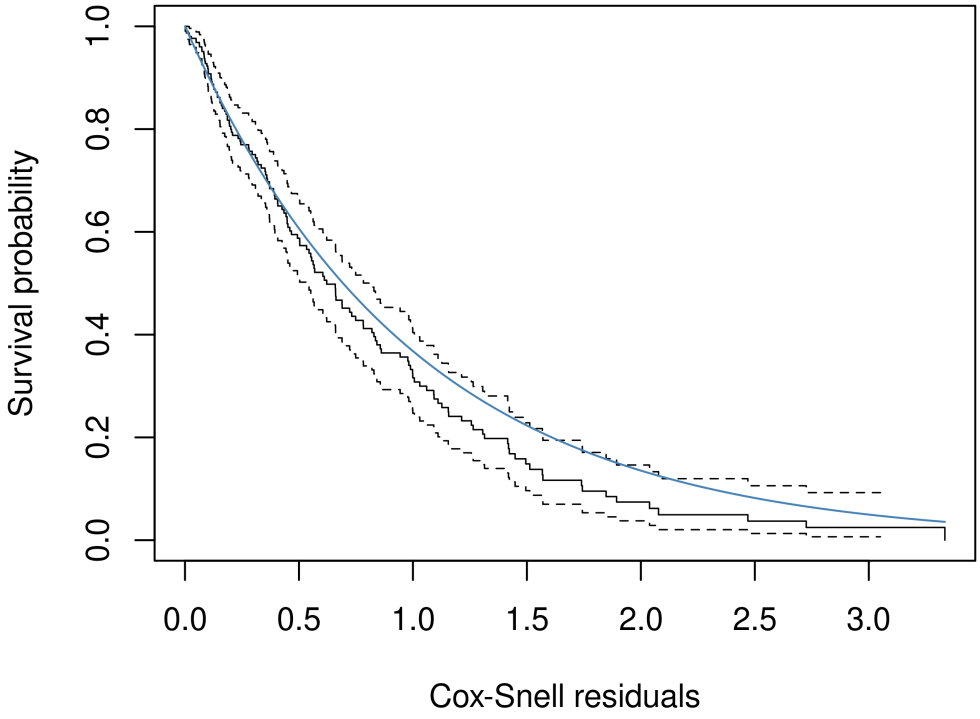


Figure C.29: Survival function of the Cox-Snell residuals obtained from the 'final' joint model fit to the PBC data in Section 7.5. The steel-blue overlaid curve represents the theoretical unit exponential distribution.

C.3.7 Quantile-quantile plot of residuals from final model

QQ plots for the Pearson residuals from the multivariate joint model in Section 7.5 across each biomarker are presented in Figure C.30

C.3.8 Quantile-quantile plot of random effects from final model

QQ plots for the estimated random effects from the multivariate joint model in Section 7.5, specified in (7.4), are shown in Figure C.31

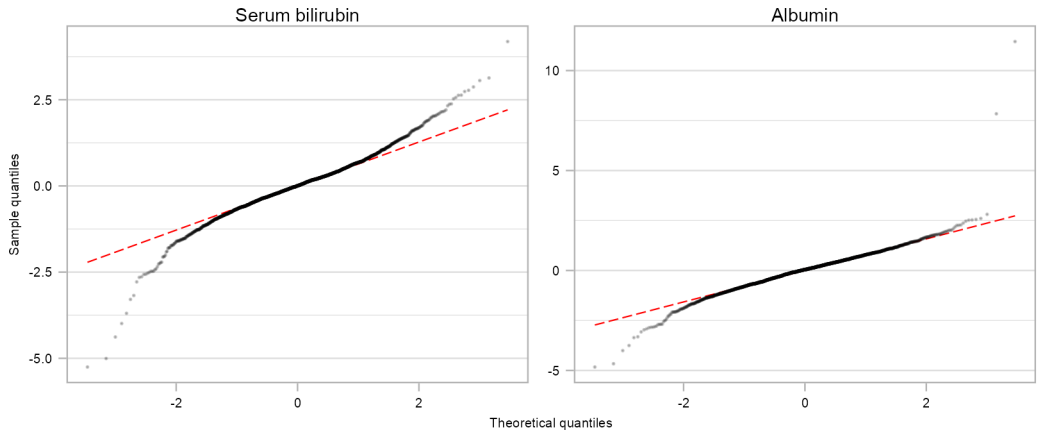


Figure C.30: Normal quantile-quantile plots for the Pearson residuals obtained for each longitudinal sub-model in the 'final' joint model fit to the PBC data in Section 7.5; the biomarker is shown in the panel title.

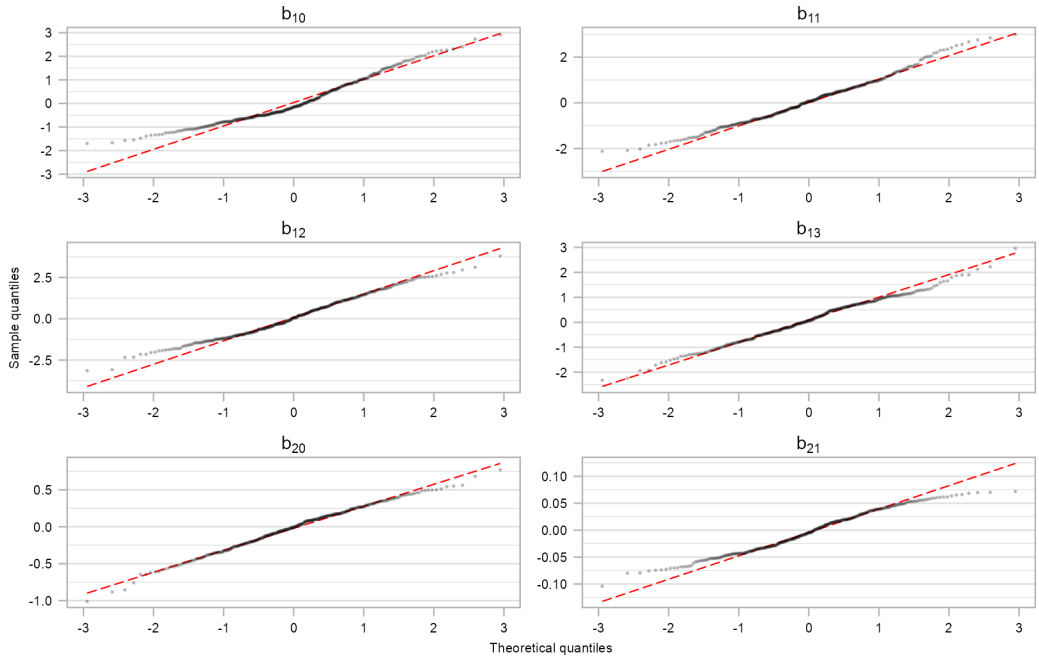


Figure C.31: Normal quantile-quantile plots for the estimated random effects from the 'final' joint model fit to the PBC data in Section 7.5; the random effect is shown in the panel title and corresponds to (7.4).

C.3.9 Posterior density of random effects from final model

Samples from the posterior distribution $f\left(\boldsymbol{b}_i|\boldsymbol{Y}_i, T_i, \Delta_i; \hat{\boldsymbol{\Omega}}\right)$ are shown in Figure C.32. 3500 samples are drawn post 500 samples of burn-in and the Metropolis-Hastings acceptance rate controlled to be $\approx 20\%$, this then repeated across $i = 1, \ldots, n$ subjects. The red

dashed line is the theoretical normal density $N\big(0, \hat{\mathrm{D}}_{xx}\big)$ with $\hat{\mathrm{D}}_{xx}$ the corresponding estimated variance component.

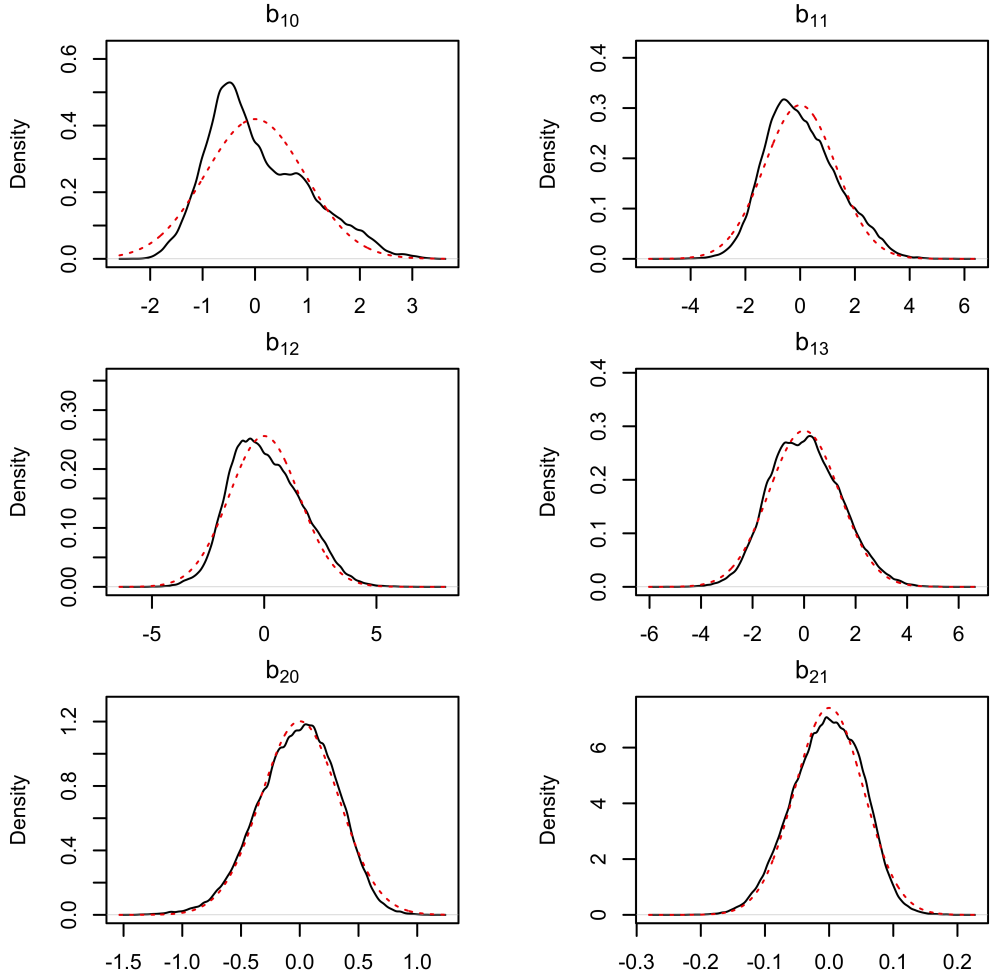


Figure C.32: Posterior densities of the estimated random effects from the 'final' joint model fit to the PBC data in Section 7.5; the random effect is shown in the panel title and corresponds to (7.4).

Appendix D

The R package gmvjoint

The aim of the work presented (particularly in Chapter 4) lead to development of the R package gmvjoint, which is available on CRAN. Since the package was used extensively in generation of all presented results, a brief overview of all 'major' functions is given in this short Appendix.
All source code is available at https://github.com/jamesmurray7/gmvjoint.

D.1 Data generation: simData

The facility to generate data under a joint model is vitally important to validation of the proposed method, and we simulate copious amounts of said data in the work presented in the main body. The bespoke function simData allows us to simulate under a wide variety of scenarios which may be of interest.

Underlying features of the data such as the sample size, the maximal number of follow-up measurements (i.e. r in Section 3.4.1), the length of follow-up (κ in the same section) and whether or not the follow-up times are regularly spaced or randomly drawn from Unif$[0, \kappa]$ can be controlled from the call to simData by arguments n; ntms; fup; and regular.times, respectively.

For parameter values, the true fixed effects $\boldsymbol{\beta}$ are supplied as a $K \times 4$ *matrix*, wherein each row contains $\boldsymbol{\beta}_k$, providing the true fixed effect for the intercept, time, standard normal realization and random Bernoulli draw for the k^{th} response. The variance-covariance matrix on random effects D is supplied as a matrix, which is internally checked for positive semi-definiteness. The degrees of freedom of the random effects can be supplied by the dof argument. The conditional distribution of the response $\boldsymbol{Y}_{ik}|\boldsymbol{b}_{ik}\ \forall\ i = 1, \ldots, n$ is provided as a list of length K and can take values "gaussian"; "poisson"; "binomial"; "Gamma";

"negbin"; and "genpois".

The random effects specification for each argument can be set by argument `random.formula`, which takes a `list` of `formula` objects (only random intercept, and intercept-and-slope are implemented, though). Dispersion parameters $\boldsymbol{\sigma}$ are provided as a `list` of length K, where each element corresponds to the dispersion model explained in Section 4.5 and outlined for each family in Table 4.1; time-varying specifications can be supplied by simply providing a vector in the k^{th} position and corresponding item in `disp.formulas` set to an appropriate `formula` object.

Finally, survival times are generated by the methodology outlined in Sections 2.5.2 and 2.5.3. A K-vector of parameters `gamma` specifies the association held by the random effects for each response, and time-invariant survival parameters are provided as a vector of length 2 `zeta`, corresponding to the standard normal realization and Bernoulli draw (i.e. the baseline covariates). The baseline hazard is additionally controlled by argument `theta`.

D.2 Workhorse function: `joint`

The main workhorse function is `joint`. This takes in a few key arguments:

`long.formulas`: A `list` containing the K `formula`s constructing the GLMM sub-models in desired joint model. These must match the syntax expected by `glmmTMB` (Brooks et al., 2017):

```
<response> ~ <fixed effect specification> + (<random effect
   specification | <grouping variable>)
```

`surv.formula`: A `formula` object usable by `coxph` from the `survival` package (Therneau, 2015):

```
Surv(<survival time>, <event indicator>) ~ <survival covariates>
```

`data`: A `data.frame` object containing all necessary variables.
`family`: A `list` of length K, with each element a character string matching one of "gaussian"; "poisson"; "binomial"; "Gamma"; "negbin"; or "genpois". The k^{th} element of `family` 'matches up' with the corresponding k^{th} `long.formulas` entry.
`disp.formula`: A `list` of length K, each element specifying the dispersion `formula` required for each sub-model. Elements are ignored if corresponding family doesn't model dispersion; if left untouched then intercept-only dispersion is modelled.
`control`: A `list` specifying control arguments. This allows for (amongst other things) changing of convergence criterion outlined in Section 3.2.3; the step-size in numerical dif-

ferentiation routines (e.g. central differencing in Appendix A.1); and increasing/decreasing the number of quadrature nodes.

The `joint` function then sends these arguments through many internal functions to obtain initial conditions; execute an EM algorithm until convergence; and then undergo post-processing before finally returning an object with class `joint`, which we give an outline of in the next section.

D.3 R object of class joint and its S3 methods

Once the `joint` object has been successfully fit, it returns an `R` object with class `joint`. Within `R`, a style of pseudo-object oriented programming functions called 'S3 methods'[1] are available; these work essentially by inspecting the class on a given object when called using a reserved generic function name, e.g. `print`, and then executing a specific function written for said specific class `joint`.

The user can `print` a `joint` object to e.g. the `R` console window by simply typing its object name, where the user then observes association parameter estimates, along with information about the model and the data. A more in-depth representation of the joint model is provided by the `summary` S3 method, which prints, amongst other items, *all* parameter estimates, their standard errors and p-values. LaTeX-ready tables are provided by the `xtable` S3 method. The user can extract specific 'parts' of the joint model such as: The fixed effects by `fixef`; random effects by `ranef`; variance-covariance matrix by `vcov`; and the log-likelihood and AIC by `logLik` and `AIC`, respectively.

Toy exemplar usage of `gmvjoint` which demonstrates data generation, model fitting, and these S3 methods is provided in the next section.

D.4 Limitations of gmvjoint

`gmvjoint` is very much fledgling software developed to facilitate and transportability of the novel methodologies introduced. With this being said, there are a few limitations with the software, here we list those the interested reader may wish to take into consideration:

- Owing to the steps taken in generation of design matrices, the survival term can not contain interactions.
- The time variable and subject identifier **must** be called `time` and `id` in the `data` provided by the user.

[1] Named as such as they were first introduced in version 3 of `R`'s predecessor, `S`.

- Data must be balanced (i.e. no missing values), else `joint` throws an error once the underlying `C++` routines are called.
- Models for the conditional expectation of the response, as well as dispersion models, are not flexibly specified (see Table 4.1 in Chapter 4), whereas one would usually be able to specify link functions herein e.g. `family='binomial'(link='log')`.

D.5 Example use of `gmvjoint`

In the upcoming example, `typewriter font with a grey background` denotes code which is executed in `R`, with `typewriter font with no background` the output. We begin by simulating some data matching the simulation study specification carried out in Section 4.5.3

```
library(gmvjoint)
set.seed(1995)
# Default values as outlined in Section 4.5.2
D <- diag(c(0.25, 0.05, 0.50, 0.09, 2.00))
D[1,3] <- D[3,1] <- D[1,5] <- D[5,1] <- D[3,5] <- D[5,3] <- 0.125
sim.dat <- simData(
  beta = rbind(
    c(-2, .1,-.1,.2), # Attached to Gaussian response
    c( 2,-.1, .1,.2), # Poisson
    c( 1, -1,  1,-1)  # Binomial
  ), D = D,
  # Survival parameters (zeta is attached to the binary x_2 only)
  gamma = c(0.5, -0.5, 0.5), zeta = c(0, -0.2),
  ntms = 10L, n = 500L,
  family = list("gaussian", "poisson", "binomial"),
  # Residual variance for Gaussian term
  sigma = list(0.16, 0, 0),
  # Random intercept only for the binary response
  random.formulas = list(~time, ~time, ~1)
)
# This generates a list of the "full" longitudinal and "survival"
   data.
# Taking the former forwards, simulated data preview...
sim.dat <- sim.dat$data
head(sim.dat)
```

```
   id      time       cont bin        Y.1 Y.2 Y.3  survtime status
1   1 0.0000000  1.0607633   1 -2.1009485   7   0 2.5608142      1
2   1 0.5555556  1.0607633   1 -1.5464562  10   0 2.5608142      1
```

```
3   1 1.1111111  1.0607633   1 -2.2722669   8   0 2.5608142      1
4   1 1.6666667  1.0607633   1 -0.9587497   5   0 2.5608142      1
5   1 2.2222222  1.0607633   1 -1.0971355   6   0 2.5608142      1
11  2 0.0000000 -0.3355017   1 -2.5143327  11   0 0.9009338      1
```

Next, we define formula objects we use to fit the joint model, and fit the model using the workhorse function `joint`. Formulae are defined outside of the call to `joint` for readability, although they could equally be simply defined 'within' this function call.

```
longs <- list(
  Y.1 ~ time + cont + bin + (1 + time|id),
  Y.2 ~ time + cont + bin + (1 + time|id),
  Y.3 ~ time + cont + bin + (1|id)
)
surv <- Surv(survtime, status) ~ bin
fit <- joint(
  long.formulas = longs, surv.formula = surv,
  data = sim.dat, family = list("gaussian", "poisson", "binomial")
)
# S3 method for printing
fit
```

```
Number of subjects: 500
Number of events: 90 (18.00%)

====================
Model specification
====================
Multivariate longitudinal process specifications:
Y.1 (gaussian): ~ time + cont + bin + (1 + time | id)
Y.2 (poisson): ~ time + cont + bin + (1 + time | id)
Y.3 (binomial): ~ time + cont + bin + (1 | id)

Survival sub-model specification:
Surv(survtime, status) ~ bin

Association parameter estimates:
       Y.1        Y.2        Y.3
 0.5022092 -0.4497374  0.3773821
```

With the fitted joint model, we can then enact several S3 methods as alluded to in the previous section. These are presented across several code-output 'chunks' which follow.

```
# S3 method for model summary
summary(fit)
```

```
Number of subjects: 500
Number of events: 90 (18.00%)
Number of responses: 3; dimension of random effects: 5
Median [IQR] profile length: 10 [9, 10]

Model fit statistics ----
  log.Lik       AIC       BIC
-17153.14  34370.27  34609.33
Degrees of freedom: 32

Longitudinal processes ----
Y.1 (Gaussian)
Call:
Y.1 ~ time + cont + bin + (1 + time | id)
                              SE       Z p-value   2.5%  97.5%
Y.1_(Intercept) -2.01205420 0.040 -49.885   0.000 -2.091 -1.933
Y.1_time         0.12012154 0.012   9.774   0.000  0.096  0.144
Y.1_cont        -0.08178707 0.027  -3.044   0.002 -0.134 -0.029
Y.1_bin          0.22080319 0.053   4.184   0.000  0.117  0.324
sigma^2_1        0.15537859 0.004  38.110   0.000  0.147  0.163

Y.2 (Poisson)
Call:
Y.2 ~ time + cont + bin + (1 + time | id)
                              SE      Z p-value   2.5%  97.5%
Y.2_(Intercept)  2.07276941 0.045 45.933    0.00  1.984  2.161
Y.2_time        -0.09997427 0.015 -6.741    0.00 -0.129 -0.071
Y.2_cont         0.12526977 0.034  3.679    0.00  0.059  0.192
Y.2_bin          0.12348258 0.066  1.882    0.06 -0.005  0.252

Y.3 (Binomial)
Call:
Y.3 ~ time + cont + bin + (1 | id)
                             SE       Z p-value   2.5%  97.5%
Y.3_(Intercept)  1.109736 0.133   8.368       0  0.850  1.370
Y.3_time        -1.024679 0.046 -22.045       0 -1.116 -0.934
Y.3_cont         1.021418 0.088  11.609       0  0.849  1.194
Y.3_bin         -1.377318 0.155  -8.909       0 -1.680 -1.074

Event-time sub-model: ----
Call: Surv(survtime, status) ~ bin
                        SE      Z p-value   2.5%  97.5%
zeta_bin  -0.2334932 0.236 -0.991   0.322 -0.695  0.228
```

```
gamma_Y.1  0.5022092 0.117  4.302   0.000  0.273  0.731
gamma_Y.2 -0.4497374 0.106 -4.249   0.000 -0.657 -0.242
gamma_Y.3  0.3773821 0.098  3.838   0.000  0.185  0.570

Computation summary: ----
Number of EM iterations: 7,
Time spent in EM algorithm: 5.89s
Total computation time: 10.34s.
```

```
# Fixed effects (longitudinal process, the default)
fixef(fit)
# and the survival
fixef(fit, what = "surv")
```

```
Y.1_(Intercept)        Y.1_time        Y.1_cont         Y.1_bin Y.2_(Intercept)        Y.2_time
    -2.01205420      0.12012154     -0.08178707      0.22080319      2.07276941     -0.09997427
       Y.3_time        Y.3_cont         Y.3_bin
    -1.02467854      1.02141792     -1.37731756

  zeta_bin    gamma_1    gamma_2    gamma_3
-0.2334932  0.5022092 -0.4497374  0.3773821
```

```
# Random effects estimates (first five subjects only)
ranef(fit)[1:5,]
```

```
     Y.1_(Intercept)     Y.1_time Y.2_(Intercept)    Y.2_time Y.3_(Intercept)
[1,]     -0.05814563  0.199440189      -0.1672700 -0.08070501      -1.2257325
[2,]     -0.44714415 -0.007098271      -0.0673282 -0.13425896      -0.3250280
[3,]     -1.13464512  0.240416941      -1.2778383 -0.11237033      -0.1985621
[4,]     -0.46574655 -0.255950104      -0.8261767 -0.08962955       2.3944808
[5,]      0.72637241 -0.113025740       0.7535331 -0.07375549      -2.0905239
```

```
# Model log-likelihood and AIC
logLik(fit)
AIC(fit)
```

```
'log Lik.' -17153.14 (df=32)
34370.27
```

D.6 Usage of C++ to reduce computation time

In Section 3.7 we noted that packages which allow for integration of C++ in R are used to overcome computational bottlenecks. The interfacing between R and compiled languages

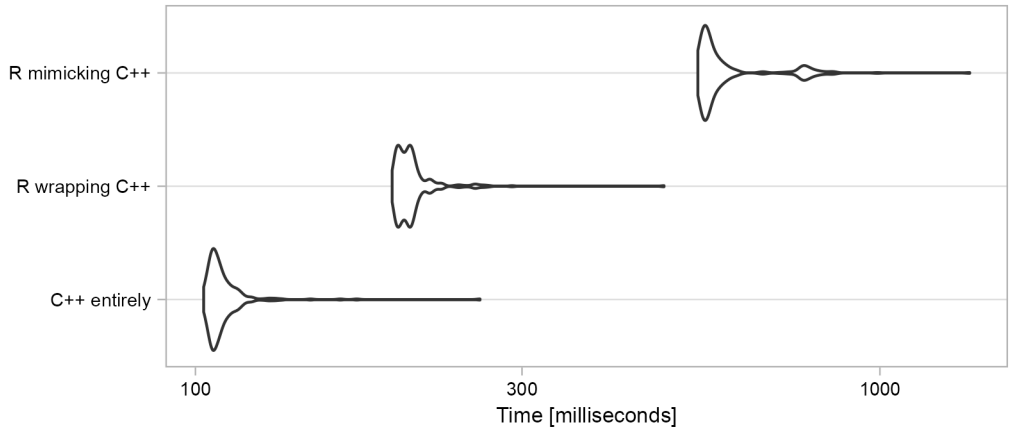


Figure D.1: Violin plot of elapsed times taken for 250 benchmarked replicates of updates to the survival parameter vector $\hat{\boldsymbol{\Phi}}$. A log scale is employed on the x-axis between ticks.

(such as `C++`) is becoming the norm, with an estimated $\approx 13\%$ of all packages on `CRAN` using `Rcpp` (Eddelbuettel and François, 2011).

Briefly in this section we provide an example of the performance gains with respect to computation time we observed in development of the methods in Chapters 3 and 4. The update for survival parameters is a known bottleneck (Hickey et al., 2018a): The conditional expectation (3.14) needs to be numerically differentiated n times at each EM iteration.

We consider three different strategies to obtaining the updated survival parameters $\hat{\boldsymbol{\Phi}}$ (2.12). We arbitrarily chose the univariate joint model we fit to (log) serum bilirubin in the PBC application in Section 7.4.1. The first method implements the objective conditional expectation (3.14), along with forward and central differencing routines – as outlined in Appendix A.1 – all in `C++`; the next implements the conditional expectation in `C++`, but calls the `R` package `pracma` (Borchers, 2022) to carry out this numerical differentiation. Lastly, we consider an implementation written in `R` entirely.

The `R` package `microbenchmark` (Mersmann, 2019) is used to replicate each strategy 250 times and benchmark the elapsed times against one another; a violin plot of elapsed times is given in Figure D.1. Here we notice that simply rewriting the conditional expectation in `C++` corresponds to a near three-fold decrease in average computation time (602.96 $\rightarrow$ 208.09ms), and further implementing the numerical differentiation into `C++` approximately a 2-fold decrease (208.09 $\rightarrow$ 110.17ms). Although these gains may appear small, one must consider that these will all only increase with n. In addition, this benchmarking solely considered 'one iteration' and in reality these performance gains will 'add up' over the full implementation of the iterative EM algorithm.

Bibliography

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19, 716–723. doi:10.1109/TAC.1974.1100705.

Alam, K., Maity, A., Sinha, S.K., Rizopoulos, D., Sattar, A., 2021. Joint modeling of longitudinal continuous, longitudinal ordinal, and time-to-event outcomes. Lifetime Data Analysis 27, 64–90. doi:10.1007/s10985-020-09511-3.

Alsefri, M., Sudell, M., García-Fiñana, M., Kolamunnage-Dona, R., 2020. Bayesian joint modelling of longitudinal and time to event data: A methodological review. BMC Medical Research Methodology 20, 94. doi:10.1186/s12874-020-00976-2.

Andersen, P.K., Gill, R.D., 1982. Cox's regression model for counting processes: A large sample study. The Annals of Statistics 10, 1100–1120.

Andrinopoulou, E.R., Clancy, J.P., Szczesniak, R., 2020. Multivariate joint modeling to identify markers of growth and lung function decline that predict cystic fibrosis pulmonary exacerbation onset. BMC pulmonary medicine 20, 1–11. doi:10.1186/s12890-020-1177-z.

Andrinopoulou, E.R., Harhay, M.O., Ratcliffe, S.J., Rizopoulos, D., 2021. Reflection on modern methods: Dynamic prediction using joint models of longitudinal and time-to-event data. International Journal of Epidemiology 50, 1731–1743. doi:10.1093/ije/dyab047.

Andrinopoulou, E.R., Rizopoulos, D., 2016. Bayesian shrinkage approach for a joint model of longitudinal and survival outcomes assuming different association structures. Statistics in Medicine 35, 4813–4823. doi:10.1002/sim.7027.

Austin, P.C., 2012. Generating survival times to simulate Cox proportional hazards models with time-varying covariates. Statistics in Medicine 31, 3946–3958.

Baghishani, H., Mohammadzadeh, M., 2012. Asymptotic normality of posterior distributions for generalized linear mixed models. Journal of Multivariate Analysis 111, 66–77. doi:10.1016/j.jmva.2012.05.003.

Barrett, J., Su, L., 2017. Dynamic predictions using flexible joint models of longitudinal and time-to-event data. Statistics in medicine 36, 1447–1460. doi:10.1002/sim.7209.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67, 1–48. doi:10.18637/jss.v067.i01.

Bender, R., Augustin, T., Blettner, M., 2005. Generating survival times to simulate Cox proportional hazards models. Statist. Med. 24, 1713–1723.

Bernhardt, P.W., Zhang, D., Wang, H.J., 2015. 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. doi:10.1016/j.csda.2014.11.011.

Berzuini, C., Larizza, C., 1996. A unified approach for modeling longitudinal and failure time data, with application in medical monitoring. IEEE Transactions on Pattern Analysis and Machine Intelligence 18, 109–123. doi:10.1109/34.481537.

Beuers, U., Gershwin, M.E., Gish, R.G., Invernizzi, P., Jones, D.E., Lindor, K., Ma, X., Mackay, I.R., Parés, A., Tanaka, A., Vierling, J.M., Poupon, R., 2015. Changing nomenclature for PBC: From 'cirrhosis' to 'cholangitis'. Clinics and Research in Hepatology and Gastroenterology 39, e57–e59. doi:10.1016/j.clinre.2015.08.001.

Bolker, B., 2022. GLMM FAQ. https://bbolker.github.io/mixedmodels-misc/glmmFAQ. Accessed: June 20, 2023.

Borchers, H.W., 2022. pracma: Practical Numerical Math Functions. URL: https://CRAN.R-project.org/package=pracma. r package version 2.4.2.

Brooks, M.E., Kristensen, K., van Benthem, K.J., Magnusson, A., Berg, C.W., Nielsen, A., Skaug, H.J., Maechler, M., Bolker, B.M., 2017. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. The R Journal 9, 378–400. URL: https://journal.r-project.org/archive/2017/RJ-2017-066/index.html.

Bycott, P., Taylor, J., 1998. A comparison of smoothing techniques for CD4 data measured with error in a time-dependent Cox proportional hazards model. Statistics in medicine 17, 2061–2077.

Chen, L.M., Ibrahim, J.G., Chu, H., 2011. Sample size and power determination in joint modeling of longitudinal and survival data. Statistics in Medicine 30, 2295–2309. doi:10.1002/sim.4263.

Chi, Y.Y., Ibrahim, J.G., 2005. Joint Models for Multivariate Longitudinal and Multivariate Survival Data. Biometrics 62, 432–445. doi:10.1111/j.1541-0420.2005.00448.x.

Choi, J., Cai, J., Zeng, D., Olshan, A.F., 2015. Joint analysis of survival time and longitudinal categorical outcomes. Statistics in biosciences 7, 19–47. doi:10.1007/s12561-013-9091-z.

Conway, R.W., Maxwell, W.L., 1962. A queuing model with state dependent service rates. Journal of Industrial Engineering 12, 132–136.

Cox, D.R., 1972. Regression models and life-tables. Journal of the Royal Statistical Society. Series B (Methodological) 34, 187–220. doi:10.1111/j.2517-6161.1972.tb00899.x.

Cox, D.R., Snell, E.J., 1968. A general definition of residuals. Journal of the Royal Statistical Society. Series B (Methodological) 30, 248–275. URL: http://www.jstor.org/stable/2984505.

Crowther, M.J., Abrams, K.R., Lambert, P.C., 2013. Joint modeling of longitudinal and survival data. Stata Journal 13, 165–184(20).

Crowther, M.J., Andersson, T.M.L., Lambert, P.C., Abrams, K.R., Humphreys, K., 2016. Joint modelling of longitudinal and survival data: Incorporating delayed entry and an assessment of model misspecification. Statistics in Medicine 35, 1193–1209. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6779, doi:https://doi.org/10.1002/sim.6779.

Dafni, U.G., Tsiatis, A.A., 1998. Evaluating surrogate markers of clinical outcome when measured with error. Biometrics 54, 1445–1462. doi:10.2307/2533670.

Dai, H., Pan, J., 2018. Joint modelling of survival and longitudinal data with informative observation times: Joint modelling. Scandinavian Journal of Statistics 45. doi:10.1111/sjos.12314.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39, 1–38. doi:10.1111/j.2517-6161.1977.tb01600.x.

Dil, E., Karasoy, D., 2020. Joint modeling of a longitudinal measurement and parametric survival data with application to primary biliary cirrhosis study. Pakistan Journal of Statistics and Operation Research 16, 295–304. doi:10.18187/pjsor.v16i2.3131.

Eddelbuettel, D., François, R., 2011. Rcpp: Seamless R and C++ integration. Journal of Statistical Software 40, 1–18. URL: https://www.jstatsoft.org/v40/i08/, doi:10.18637/jss.v040.i08.

Eddelbuettel, D., Sanderson, C., 2014. (rcpparmadillo): Accelerating R with high-performance C++ linear algebra. Computational Statistics and Data Analysis 71, 1054–1063. URL: http://dx.doi.org/10.1016/j.csda.2013.02.005.

Faucett, C.L., Thomas, D.C., 1996. Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. Statistics in medicine 15, 1663–1685. doi:10.1002/(SICI)1097-0258(19960815)15:15<1663::AID-SIM294>3.0.CO;2-1.

Fournier, D.A., Skaug, H.J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M.N., Nielsen, A., Sibert, J., 2012. AD Model Builder: Using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. Optimization Methods and Software 27, 233–249. doi:10.1080/10556788.2011.597854.

Fox, J., Weisberg, S., 2019. An R Companion to Applied Regression. Third ed., Sage, Thousand Oaks CA. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Friendly, M., Monette, G., Fox, J., 2013. Elliptical Insights: Understanding Statistical Methods through Elliptical Geometry. Statistical Science 28, 1–39. doi:10.1214/12-STS402.

Furgal, A.K., Sen, A., Taylor, J.M., 2019. Review and comparison of computational approaches for joint longitudinal and time-to-event models. International Statistical Review 87, 393–418. doi:10.1111/insr.12322.

Garcia-Hernandez, A., Rizopoulos, D., 2018. %JM: A SAS macro to fit jointly generalized mixed models for longitudinal data and time-to-event responses. Journal of Statistical Software 84, 1–29. doi:10.18637/jss.v084.i12.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Hothorn, T., 2021. mvtnorm: Multivariate Normal and t Distributions. URL: https://CRAN.R-project.org/package=mvtnorm. r package version 1.1-3.

Gould, L.A., Boye, M.E., Crowther, M.J., Ibrahim, J.G., Quartey, G., Micallef, S., Bois, F.Y., 2015. Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA bayesian joint modeling working group. Statistics in Medicine 34, 2181–2195. doi:10.1002/sim.6141.

Gruttola, V.D., Tu, X.M., 1994. Modelling progression of CD4-lymphocyte count and its relationship to survival time. Biometrics 50, 1003–1014.

Hand, D., Christen, P., 2018. A note on using the F-measure for evaluating record linkage algorithms. Statistics and Computing 28, 539–547. doi:10.1007/s11222-017-9746-6.

Henderson, R., Diggle, P., Dobson, A., 2000. Joint modelling of longitudinal measurements and event time data. Biostatistics 4, 465–480. doi:10.1093/biostatistics/1.4.465.

Henderson, R., Diggle, P., Dobson, A., 2002. Identification and efficacy of longitudinal markers for survival. Biostatistics 3, 33–50. doi:10.1093/biostatistics/3.1.33.

Hickey, G.L., Philipson, P., Jorgensen, A., Kolamunnage-Dona, R., 2016. Joint modelling of time-to-event and multivariate longitudinal outcomes: Recent developments and issues. BMC Med. Res. Methodol. 117. doi:10.1186/s12874-016-0212-5.

Hickey, G.L., Philipson, P., Jorgensen, A., Kolamunnage-Dona, R., 2018a. joineRML: A joint model and software package for time-to-event and multivariate longitudinal outcomes. BMC Med. Res. Methodol. 50. doi:10.1186/s12874-018-0502-1.

Hickey, G.L., Philipson, P., Jorgensen, A., Kolamunnage-Dona, R., 2018b. Joint models of longitudinal and time-to-event data with more than one event time outcome: A review. The International Journal of Biostatistics 14, 20170047. doi:10.1515/ijb-2017-0047.

Horrocks, J., van Den Heuvel, M.J., 2009. Prediction of pregnancy: A joint model for longitudinal and binary data. Bayesian Analysis 4, 523–538. doi:10.1214/09-BA419.

Hsieh, F., Tseng, Y.K., Wang, J.L., 2006. Joint modeling of survival and longitudinal data: Likelihood approach revisited. Biometrics 62, 1037–1043. doi:/10.1111/j.1541-0420.2006.00570.x.

Huang, A., 2017. Mean-parametrized Conway–Maxwell–Poisson regression models for dispersed counts. Statistical Modelling 17, 359–380. doi:10.1177/1471082X17697749.

Hwang, Y.T., Tsai, H.Y., Chang, Y.J., Kuo, H.C., Wang, C.C., 2011. The joint model of the logistic model and linear random effect model — An application to predict orthostatic hypertension for subacute stroke patients. Computational Statistics & Data Analysis 55, 914–923. doi:10.1016/j.csda.2010.07.024.

Hwang, Y.T., Wang, C.C., Wang, C.H., Tseng, Y.K., Chang, Y.J., 2015. Joint model of multiple longitudinal measures and a binary outcome: An application to predict orthostatic hypertension for subacute stroke patients. Biometrical Journal 57, 661–675. doi:10.1002/bimj.201400044.

Ibrahim, J.G., Chu, H., Chen, L.M., 2010. Basic concepts and methods for joint models of longitudinal and survival data. Journal of Clinical Oncology 28, 2796–2801. doi:10.1200/JCO.2009.25.0654.

Ismail, N., Jemain, A.A., 2007. Handling overdispersion with negative binomial and generalized Poisson regression models, in: Casualty Actuarial Society Forum, Winter 2007, pp. 103–149.

Kalbfleisch, J.D., Prentice, R.L., 2002. The Statistical Analysis of Failure time data. 2nd ed., John Wiley & Sons.

Khan, S.A., Basharat, N., 2022. Accelerated failure time models for recurrent event data analysis and joint modeling. Computational Statistics 37, 1569–1597. URL: https://doi.org/10.1007/s00180-021-01171-7, doi:10.1007/s00180-021-01171-7.

Kristensen, K., Nielsen, A., Berg, C.W., Skaug, H., Bell, B.M., 2016. TMB: Automatic differentiation and Laplace approximation. Journal of Statistical Software 70, 1—21. doi:10.18637/jss.v070.i05.

Kubinec, R., 2023. Ordered Beta regression: A parsimonious, well-fitting model for continuous data with lower and upper bounds. Political Analysis 31, 519—536. doi:10.1017/pan.2022.20.

Laird, N.M., Ware, J.H., 1982. Random-effects models for longitudinal data. Biometrics 38, 963–974. doi:10.2307/2529876.

Lemieux, L., 2009. Monte Carlo and quasi-Monte Carlo sampling. Springer.

Li, C., Xiao, L., Luo, S., 2021. Joint model for survival and multivariate sparse functional data with application to a study of Alzheimer's disease. Biometrics doi:10.1111/biom.13427.

Li, D., Wang, X., Song, S., Zhang, N., Dey, D.K., 2015. Flexible link functions in a joint model of binary and longitudinal data. Stat 4, 320–330. doi:10.1002/sta4.98.

Li, K., Luo, S., 2017. Functional joint model for longitudinal and time-to-event data: An application to Alzheimer's disease. Statistics in medicine 36, 3560–3572. doi:10.1002/sim.7381.

Li, N., Elashoff, R.M., Li, G., Saver, J., 2010. Joint modeling of longitudinal ordinal data and competing risks survival times and analysis of the NINDS rt-PA stroke trial. Statistics in medicine 29, 546–557. doi:10.1002/sim.3798.

Li, S., 2022. FastJM: Semi-Parametric Joint Modeling of Longitudinal and Survival Data. URL: https://CRAN.R-project.org/package=FastJM. r package version 1.2.0.

Li, S., Li, N., Wang, H., Zhou, J., Zhou, H., Li, G., 2022. Efficient algorithms and implementation of a semiparametric joint model for longitudinal and competing risk

data: With applications to massive Biobank data. Computational and Mathematical Methods in Medicine 2022, 1362913. doi:10.1155/2022/1362913.

Lin, H., McCulloch, C.E., Mayne, S.T., 2002. Maximum likelihood estimation in the joint analysis of time-to-event and multiple longitudinal variables. Statistics in Medicine 21, 2369–2382. doi:10.1002/sim.1179.

Long, J.D., Mills, J.A., 2018. Joint modeling of multivariate longitudinal data and survival data in several observational studies of Huntington's disease. BMC medical research methodology 18, 1–15. doi:10.1186/s12874-018-0592-9.

Martins, R., 2022. A flexible link for joint modelling longitudinal and survival data accounting for individual longitudinal heterogeneity. Statistical Methods & Applications 31, 41–61. doi:10.1007/s10260-021-00566-6.

McCullagh, P., Nelder, J., 1989. Generalized Linear Models. 2nd ed., Chapman and Hall. doi:10.1201/9780203753736.

McFetridge, L.M., Asar, O., Wallin, J., 2021. Robust joint modelling of longitudinal and survival data: Incorporating a time-varying degrees-of-freedom parameter. Biometrical Journal 63, 1587–1606. doi:10.1002/bimj.202000253.

McLachlan, G.J., Krishnan, T., 2008. The EM Algorithm and Extensions, 2nd ed. Wiley-Interscience. doi:10.1002/9780470191613.

Mersmann, O., 2019. microbenchmark: Accurate Timing Functions. URL: https://CRAN.R-project.org/package=microbenchmark. r package version 1.4-7.

Murray, J., Philipson, P., 2022. A fast approximate EM algorithm for joint models of survival and multivariate longitudinal data. Computational Statistics & Data Analysis 170, 107438. doi:10.1016/j.csda.2022.107438.

Murray, J., Philipson, P., 2023. Fast estimation for generalised multivariate joint models using an approximate EM algorithm. Computational Statistics & Data Analysis 187, 107819. doi:10.1016/j.csda.2023.107819.

Murtaugh, P.A., Dickson, E.R., Van Dam, G.M., Malinchoc, M., Grambsch, P.M., Langworthy, A.L., Gips, C.H., 1994. Primary biliary cirrhosis: Prediction of short-term survival based on repeated patient visits. Hepatology 20, 126–134. doi:10.1002/hep.1840200120.

Philipson, P., Hickey, G.L., Crowther, M.J., Kolamunnage-Dona, R., 2020. Faster Monte Carlo estimation of joint models for time-to-event and multivariate longitudinal data.

Computational Statistics & Data Analysis 151, 107–10. doi:10.1016/j.csda.2020.107010.

Philipson, P., Sousa, I., Diggle, P.J., Williamson, P., Kolamunnage-Dona, R., Henderson, R., Hickey, G.L., 2018. joineR: Joint Modelling of Repeated Measurements and Time-to-Event Data. URL: https://github.com/graemeleehickey/joineR/. r package version 1.2.5.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., R Core Team, 2021. nlme: Linear and Nonlinear Mixed Effects Models. URL: https://CRAN.R-project.org/package=nlme. r package version 3.1-152.

Pinheiro, J.C., Bates, D.M., 1995. Approximations to the log-likelihood function in the nonlinear mixed-effects model. Journal of Computational and Graphical Statistics 4, 12–35. URL: http://www.jstor.org/stable/1390625.

Prentice, R.L., 1982. Covariate measurement errors and parameter estimation in a failure time regression model. Biometrika 69, 331–342.

Purohit, T., Cappell, M.S., 2015. Primary biliary cirrhosis: Pathophysiology, clinical presentation and therapy. World journal of hepatology 7, 926–941. doi:10.4254/wjh.v7.i7.926.

R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Riley, R.D., Snell, K.I., Ensor, J., Burke, D.L., Harrell Jr, F.E., Moons, K.G., Collins, G.S., 2019. Minimum sample size for developing a multivariable prediction model: PART II – binary and time-to-event outcomes. Statistics in Medicine 38, 1276–1296. doi:10.1002/sim.7992.

Ripatti, S., Larsen, K., Palmgren, J., 2002. Maximum likelihood inference for multivariate frailty models using an automated Monte Carlo EM algorithm. Lifetime Data Analysis 8, 349–360. doi:10.1023/A:1020566821163.

Rizopoulos, D., 2010. JM: An R package for the joint modelling of longitudinal and time-to-event data. Journal of Statistical Software 35, 1–33. URL: https://doi.org/10.18637/jss.v035.i09.

Rizopoulos, D., 2011. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. Biometrics 67, 819–829. doi:10.1111/j.1541-0420.2010.01546.x.

Rizopoulos, D., 2012a. Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive Gaussian quadrature rule. Computational Statistics & Data Analysis 56, 491–501. doi:10.1016/j.csda.2011.09.007.

Rizopoulos, D., 2012b. Joint models for longitudinal and time-to-event data: With applications in R. CRC press.

Rizopoulos, D., 2016. The R package JMbayes for fitting joint models for longitudinal and time-to-event data using MCMC. Journal of Statistical Software 72, 1–45. doi:10.18637/jss.v072.i07.

Rizopoulos, D., Ghosh, P., 2011. A bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. Statistics in Medicine 30, 1366–1380. doi:10.1002/sim.4205.

Rizopoulos, D., Molenberghs, G., Lesaffre, E.M., 2017. Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. Biometrical Journal 59, 1261–1276. doi:10.1002/bimj.201600238.

Rizopoulos, D., Papageorgiou, G., Miranda Afonso, P., 2021. JMbayes2: Extended Joint Models for Longitudinal and Time-to-Event Data. URL: https://CRAN.R-project.org/package=JMbayes2. r package version 0.1-8.

Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71, 319–392. doi:10.1111/j.1467-9868.2008.00700.x.

Rustand, D., van Niekerk, J., Krainski, E.T., Rue, H., Proust-Lima, C., 2023. Fast and flexible inference for joint models of multivariate longitudinal and survival data using integrated nested Laplace approximations. Biostatistics , kxad019doi:10.1093/biostatistics/kxad019.

Schwarz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6, 461–464. URL: http://www.jstor.org/stable/2958889.

Shmueli, G., Minka, T.P., Kadane, J.B., Borle, S., Boatwright, P., 2005. A useful distribution for fitting discrete data: Revival of the Conway–Maxwell–Poisson distribution. Journal of the Royal Statistical Society: Series C (Applied Statistics) 54, 127–142. doi:10.1111/j.1467-9876.2005.00474.x.

Smyth, G.K., 2005. Numerical integration. Encyclopedia of Biostatistics , 3088–3095.

Spiegelhalter, D.J., Thomas, A., Best, N.G., 1999. WinBUGS version 1.2 User Manual. MRC Biostatistics Unit, Cambridge.

Stringer, A., Bilodeau, B., 2022. Fitting generalized linear mixed models using adaptive quadrature. arXiv:2202.07864.

Sunethra, A., Sooriyarachchi, R., 2018. A joint model for exponential survival data and Poisson count data. American Journal of Applied Mathematics and Statistics 6, 72–9. doi:10.12691/ajams-6-2-6.

Sweeting, M.J., Thompson, S.G., 2011. Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. Biometrical Journal 69, 750–763. doi:10.1002/bimj.201100052.

Therneau, T.M., 2015. A Package for Survival Analysis in S. URL: https://CRAN.R-project.org/package=survival. version 2.38.

Therneau, T.M., Grambsch, P.M., 2000. Modeling Survival Data: Extending the Cox Model. Springer, New York, NY.

Tseng, Y.K., Hsieh, F., Wang, J.L., 2005. Joint modelling of accelerated failure time and longitudinal data. Biometrika 92, 587–603. URL: https://www.jstor.org/stable/20441216, doi:10.1093/biomet/92.3.587.

Tsiatis, A.A., Davidian, M., 2004. Joint modeling of longitudinal and time-to-event data: An overview. Statistica Sinica 14, 809–834.

Tsiatis, A.A., Degruttola, V., Wulfsohn, M.S., 1995. Modeling the relationship of survival to longitudinal data measured with error. applications to survival and CD4 counts in patients with AIDS. Journal of the American Statistical Association 90, 27–37. doi:10.1080/01621459.1995.10476485.

van Smeden, M., Reitsma, J.B., Riley, R.D., Collins, G.S., Moons, K.G., 2021. Clinical prediction models: Diagnosis versus prognosis. Journal of Clinical Epidemiology 132, 142–145. URL: https://www.sciencedirect.com/science/article/pii/S0895435621000135, doi:10.1016/j.jclinepi.2021.01.009.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S. Fourth ed., Springer, New York. URL: https://www.stats.ox.ac.uk/pub/MASS4/. iSBN 0-387-95457-0.

Wang, Y., Ibrahim, J.G., Zhu, H., 2020. Partial least squares for functional joint models with applications to the Alzheimer's disease neuroimaging initiative study. Biometrics 76, 1109–1119. doi:10.1111/biom.13219.

Wang, Y., Taylor, J.M.G., 2001. Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. Journal of the American Statistical Association 96, 895–905. doi:10.1198/016214501753208591.

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. URL: https://ggplot2.tidyverse.org.

Williamson, P.R., Kolamunnage-Done, R., Philipson, P., Marson, A.G., 2008. Joint modeling of longitudinal and competing risks data. Statist. Med. 27, 6426–6438. doi:10.1002/sim.3451.

Wulfsohn, M.S., Tsiatis, A.A., 1997. A joint model for survival and longitudinal data measured with error. Biometrics 53, 330–339. doi:10.2307/2533118.

Xu, C., Baines, P., Wang, J., 2014. Standard error estimation using the EM algorithm for the joint modeling of survival and longitudinal data. Biostatistics 15, 731–44. doi:10.1093/biostatistics/kxu015.

Xu, J., Zeger, S.L., 2001. Joint analysis of longitudinal data comprising repeated measures and times to events. Journal of the Royal Statistical Society: Series C (Applied Statistics) 50, 375–387. doi:10.1111/1467-9876.00241.

Youden, W.J., 1950. Index for rating diagnostic tests. Cancer 3, 32–35. doi:10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

Zamani, H., Ismail, N., 2012. Functional form for the generalized Poisson regression model. Communications in Statistics - Theory and Methods 41, 3666–3675. doi:10.1080/03610926.2011.564742.

Zeviani, W.M., Ribeiro, P.J., Bonat, W.H., Shimakura, S.E., Muniz, J.A., 2014. The Gamma-count distribution in the analysis of experimental underdispersed data. Journal of Applied Statistics 41, 2616–2626. doi:10.1080/02664763.2014.922168.

Zhu, H., DeSantis, S.M., Luo, S., 2018. Joint modeling of longitudinal zero-inflated count and time-to-event data: A Bayesian perspective. Statistical Methods in Medical Research 27, 1258–1270. doi:10.1177/0962280216659312.