A multinomial generalized linear mixed model for clustered competing risks data

Henrique Aparecido Laureano * [1], Ricardo Rasmussen Petterle[2], Guilherme Parreira da Silva[3], and Wagner Hubo Bonat[4]

[1] Instituto de Pesquisa Pelé Pequeno Príncipe, Curitiba, Brasil

[2] Departamento de Medicina Integrada, Universidade Federal do Paraná, Curitiba, Brasil

[3,4] Laboratório de Estatística e Geoinformação, Departamento de Estatística, Universidade Federal do Paraná, Curitiba, Brasil

January 26, 2022

Abstract

Clustered competing risks data are a complex failure time data scheme. Its main characteristics are the cluster structure, which implies a latent within-cluster dependence between its elements, and its multiple variables competing to be the one responsible for the occurrence of an event, the failure. To handle this kind of data, we propose a full likelihood approach, based on a generalized linear mixed model instead a usual complex frailty model. We model the competing causes in the probability scale, in terms of the cumulative incidence function (CIF). A multinomial distribution is assumed for the competing causes and censorship, conditioned on the latent effects. The latent effects are accommodated via a multivariate Gaussian distribution. The CIF is specified as the product of an instantaneous risk level function with a failure time trajectory level function. The estimation procedure is performed through the R package TMB (Template Model Builder), an `C++` based framework with efficient Laplace approximation and automatic differentiation routines. A large simulation study is performed, based on different latent structure formulations. The model presents to be of difficult estimation, with our results converging to a latent structure where both risk and failure time trajectory levels are correlated.

*E-mail: henriqueaparecidolaureano@gmail.com

Keywords: Clustered competing risks data; Within-cluster dependence; Multinomial generalized linear mixed model (GLMM); TMB: Template Model Builder; Laplace approximation; Automatic differentiation (AD).

1 Introduction

Competing risks data, and more generally failure time data, can be modeled in two possible scales: the hazard and the probability scale, with the former being the most popular. A competing risks process can be seen as the multivariate extension of a failure time process, having multiple causes competing to be the one responsible for the desired event occurrence, properly, a failure. In Figure 1 a visual aid is provided considering m competing causes.

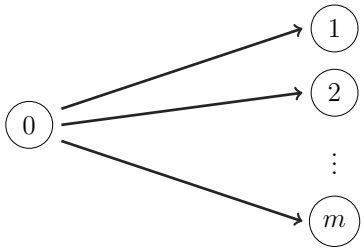

Figure 1: Illustration of competing risks process.

Failure time data is the branch of Statistics responsible to handle random variables describing the time until the occurrence of an event, a failure (Kalbfleisch and Prentice; 2002; Hougaard; 2000). The time until a failure is called survival experience, and is the modeling object. To accommodate the number of possible causes for a failure there is the competing risks data scheme. More specifically, its clustered version with groups of elements sharing some non-observed latent dependence structure.

When this framework is applied in real-world situations, we have to be able to handle with the nonoccurrence of the desired event, by any of the competing causes, for, let us say, *logistic reasons* (short-time study and outside scope causes are some examples). This, generally noninformative, nonoccurrence of the event is called censorship.

When the elements under study are organized in clusters (a family, e.g.), it opens space to what is called *family studies*. In family studies, the goal is to accommodate the non-observed latent dependence and try to understand the relationship between the family elements. In other words, how the occurrence of an event in a subject affects the survival experience for the same or similar event in its familiars.

The survival experiences is usually modeled in the hazard (failure rate) scale, and with the latent within-cluster dependence accommodation we have what is called a frailty model (Clayton; 1978; Valpel et al.; 1979; Liang et al.; 1995; Petersen; 1998). The use of frailty models implies in complicated likelihood functions and inference routines done via elaborated and slow EM algorithms (Nielsen et al.; 1992; Klein; 1992) or inefficient MCMC schemes (Hougaard; 2000). With multiple survival experiences, the general idea is the same but with even more elaborated likelihoods (Prentice et al.; 1978; Therneau and Grambsch; 2000) or mixture model approaches (Larson and Dinse; 1985; Kuk; 1992).

When in the hazard scale, the interpretations are in terms of hazard rates. A less usual scale but with a more appealing interpretation is the probability scale. For competing risks data, the work on the probability scale is done by means of the cumulative incidence function (CIF) (Andersen et al.; 2012), with the main modeling approach being the subdistribution (Fine and Gray; 1999).

For clustered competing risks data there are some available options but with a lack of predominance. The options vary in terms of likelihood specification, with its majority being designed for bivariate CIFs, where increasing the CIF's dimension is a limitation. Some of the existing options are (i) nonparametric approaches (Cheng et al.; 2007, 2009); (ii) linear transformation models (Fine; 1999; Gerds et al.; 2012); (iii) semiparametric approaches based on composite likelihoods (Shih and Albert; 2009; Cederkvist et al.; 2019), estimating equations (Scheike and Sun; 2012; Cheng and Fine; 2012), copulas (Scheike et al.; 2010), and mixtures (Naskar et al.; 2005; Shi et al.; 2013).

Besides the interpretation, by modeling the CIF it is possible to specify complex within-cluster dependence structures. We follow Cederkvist et al. (2019) and work with a CIF specification based on its decomposition in instantaneous risk and failure time trajectory functions, with both being cluster-specifics and possible correlated. As a modeling framework, we use a generalized linear mixed model (GLMM) specification. Through a GLMM we have a straightforward full likelihood specification, easy to virtually extend to any number of competing causes, and capable to allow for complex CIF structures. To make the estimation and inferential process the most efficient as possible we take advantage of state-of-art computational libraries and efficiently implemented routines under the `TMB` (Kristensen et al.; 2016) package of the R (R Core Team; 2021) statistical software.

The class of generalized linear models (GLMs) (Nelder and Wedderburn; 1972) is probably the most popular statistical modelling framework. Despite its flexibility, the GLMs are not suitable for dependent data. For the analysis of such data, Laird and Ware (1982) proposed the random effects regression models for longitudinal/repeated-measures data, and Breslow and Clayton (1993) presented the GLMMs for the analysis of non-Gaussian outcomes. In this framework, we can accommodate all competing causes of failure and censorship under a multinomial probability distribution. The latent within-cluster dependence is accommodated via a multivariate normal distribution, and the cause-specific CIFs via the model's link function.

The main goal of this work is to propose a GLMM approach to handle clustered competing risks data with a flexible within-cluster dependence structure. The model specification and the inferential routine are much simpler than the usually used approaches, increasing its practical relevance. The latent effects, the key complicator factor, are handled out by means of an efficient Laplace approximation and automatic differentiation routines. The main contributions of this article are: (i) introducing the modeling of cause/cluster-specific CIFs of clustered competing risks data into an efficient implemen-

tation of the GLMMs framework; (ii) performing a extensive simulation study to check the properties of the maximum likelihood estimator to learn the cause-specific CIF forms and the feasibility of the within-cluster dependence structure.; (iii) providing R code and `C++` implementation for the used GLMMs.

The work is organized as follows. Section 2 presents the CIF specification and the multinomial GLMM. Section 3 presents the estimation and inferential routines. Section 4 presents the performed simulation studies to check the model viability. Finally, the main contributions of the article are discussed in Section 5.

2 Model

2.1 Cluster-specific cumulative incidence function (CIF)

Consider that the observed follow-up time of a subject is given by $T = \min(T^*, C)$, where T^* denote the failure time and C denote the censoring time. Given the possible covariates $\boldsymbol{x}$, for a cause-specific of failure k, the cumulative incidence function (CIF) is defined as

$$\begin{aligned} F_k(t \mid \boldsymbol{x}) = \mathbb{P}[T \leq t,\ K = k \mid \boldsymbol{x}] &= \int_0^t f_k(z \mid \boldsymbol{x})\ \mathrm{d}z \\ &= \int_0^t \lambda_k(z \mid \boldsymbol{x})\ S(z \mid \boldsymbol{x})\ \mathrm{d}z, \quad t > 0, \quad k = 1,\ \ldots,\ K, \end{aligned}$$

where $f_k(t \mid \boldsymbol{x})$ is the (sub)density for the time to a type k failure. The subdensity is composed by the cause-specific hazard function or process $\lambda_k(t \mid \boldsymbol{x})$, representing the instantaneous rate for failures of type k at time t given $\boldsymbol{x}$ and all other failure types (competing causes). If we sum up all cause-specific hazard functions we get the overall hazard function $\lambda(t \mid \boldsymbol{x})$. From the overall hazard function we arrive in the overall survival function,

$$S(t \mid \boldsymbol{x}) = \mathbb{P}[T > t \mid \boldsymbol{x}] = \exp\left\{ - \int_0^t \lambda(z \mid \boldsymbol{x})\ \mathrm{d}z \right\},$$

the second function that compose the subdensity $f_k(t \mid \boldsymbol{x})$. A comprehensive reference for all these definitions is the book of Kalbfleisch and Prentice (2002).

To take into consideration our clustered/family structure, we follow the same CIF specification of Cederkvist et al. (2019). For two competing causes of failure, the cause-specific CIFs are specified in the following manner

$$F_k(t \mid \boldsymbol{x},\ u_1,\ u_2,\ \eta_k) = \underbrace{\pi_k(\boldsymbol{x},\ u_1,\ u_2)}_{\substack{\text{cluster-specific} \\ \text{risk level}}} \times \underbrace{\Phi[w_k g(t) - \boldsymbol{x}\boldsymbol{\gamma}_k - \eta_k]}_{\substack{\text{cluster-specific} \\ \text{failure time trajectory}}}, \quad t > 0, \quad k = 1,\ 2, \quad (1)$$

i.e., as the product of a cluster-specific risk level with a cluster-specific failure time trajectory, resulting in a cluster-specific CIF. What makes these components cluster-specific

are $\boldsymbol{u} = \{u_1,\ u_2\}$ and $\boldsymbol{\eta} = \{\eta_1,\ \eta_2\}$, Gaussian distributed latent effects with zero mean and potentially correlated,

$$\begin{bmatrix} u_1 \\ u_2 \\ \eta_1 \\ \eta_2 \end{bmatrix} \sim \begin{array}{c}\text{Multivariate} \\ \text{Normal}\end{array} \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{u_1} & \text{cov}(u_1,\ u_2) & \text{cov}(u_1,\ \eta_1) & \text{cov}(u_1,\ \eta_2) \\ & \sigma^2_{u_2} & \text{cov}(u_2,\ \eta_1) & \text{cov}(u_2,\ \eta_2) \\ & & \sigma^2_{\eta_1} & \text{cov}(\eta_1,\ \eta_2) \\ & & & \sigma^2_{\eta_2} \end{bmatrix} \right).$$

The cluster-specific risk levels are modeled by a multinomial logistic regression model with latent effects, i.e.

$$\pi_k(\boldsymbol{x}, \boldsymbol{u}) = \frac{\exp\{\boldsymbol{x}\boldsymbol{\beta}_k + u_k\}}{1 + \exp\{\boldsymbol{x}\boldsymbol{\beta}_1 + u_1\} + \exp\{\boldsymbol{x}\boldsymbol{\beta}_2 + u_2\}}, \quad k = 1,\ 2, \tag{2}$$

where the $\boldsymbol{\beta}_k$'s are the coefficients responsible for quantifying the impact of the covariates in the cause-specific risk levels. For individuals from the same cluster/family, at the same time point, the $\boldsymbol{\beta}_k$s have the well-known odds ratio interpretation.

The second component of Equation 1 is the cluster-specific failure time trajectory

$$\Phi[w_k g(t) - \boldsymbol{x}\boldsymbol{\gamma}_k - \eta_k], \quad t > 0, \quad k = 1,\ 2,$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard Gaussian distribution. With regard to the function $g(t)$, it plays a crucial role since the proposed CIF separation is only possible with it. It is used a time t transformation given by

$$g(t) = \text{arctanh}\left(\frac{t - \delta/2}{\delta/2}\right), \quad t \in (0,\ \delta), \quad g(t) \in (-\infty,\ \infty),$$

where δ depends on the data and cannot exceed the maximum observed follow-up time τ, i.e. $\delta \leq \tau$. With this Fisher-based transformation the value of the cluster-specific failure time trajectory is equal 1 at time δ. Consequently, $F_k(\delta \mid \boldsymbol{x},\ \boldsymbol{u},\ \eta_k) = \pi_k(\boldsymbol{x} \mid \boldsymbol{u})$ and we can interpret $\pi_1(\boldsymbol{x} \mid \boldsymbol{u})$ and $\pi_2(\boldsymbol{x} \mid \boldsymbol{u})$ as the cause-specific cluster-specific risk levels at time δ. The cluster-specific survival function is given by $S(t \mid \boldsymbol{x},\ \boldsymbol{u},\ \boldsymbol{\eta}) = 1 - F_1(t \mid \boldsymbol{x},\ \boldsymbol{u},\ \eta_1) - F_2(t \mid \boldsymbol{x},\ \boldsymbol{u},\ \eta_2)$.

A direct understanding of all coefficients/parameters in Equation 1 can be reached via the top-placed illustrations in Figure 2. We consider two competing causes, without covariates, and plot the cluster-specific CIF of just one failure cause. We see that:

- The β's are related to the curve's maximum value, i.e. bigger the β highest the CIF;
- The $\boldsymbol{\gamma}_k$'s are the coefficients responsible for quantifying the impact of the covariates in the cause-specific failure time trajectories, i.e. the shape of the cumulative incidence. We see that the γ's are also related with an idea of midpoint and con-

sequently, growth speed. The fact that γ_k enters negatively in the cluster-specific failure time trajectory makes that a negative value causes an advance towards the curve, whereas a positive value causes a delay;

- Last but not least, the w's. With negative values we have a decreasing curve and with positive values an increasing curve, the behavior of interest.

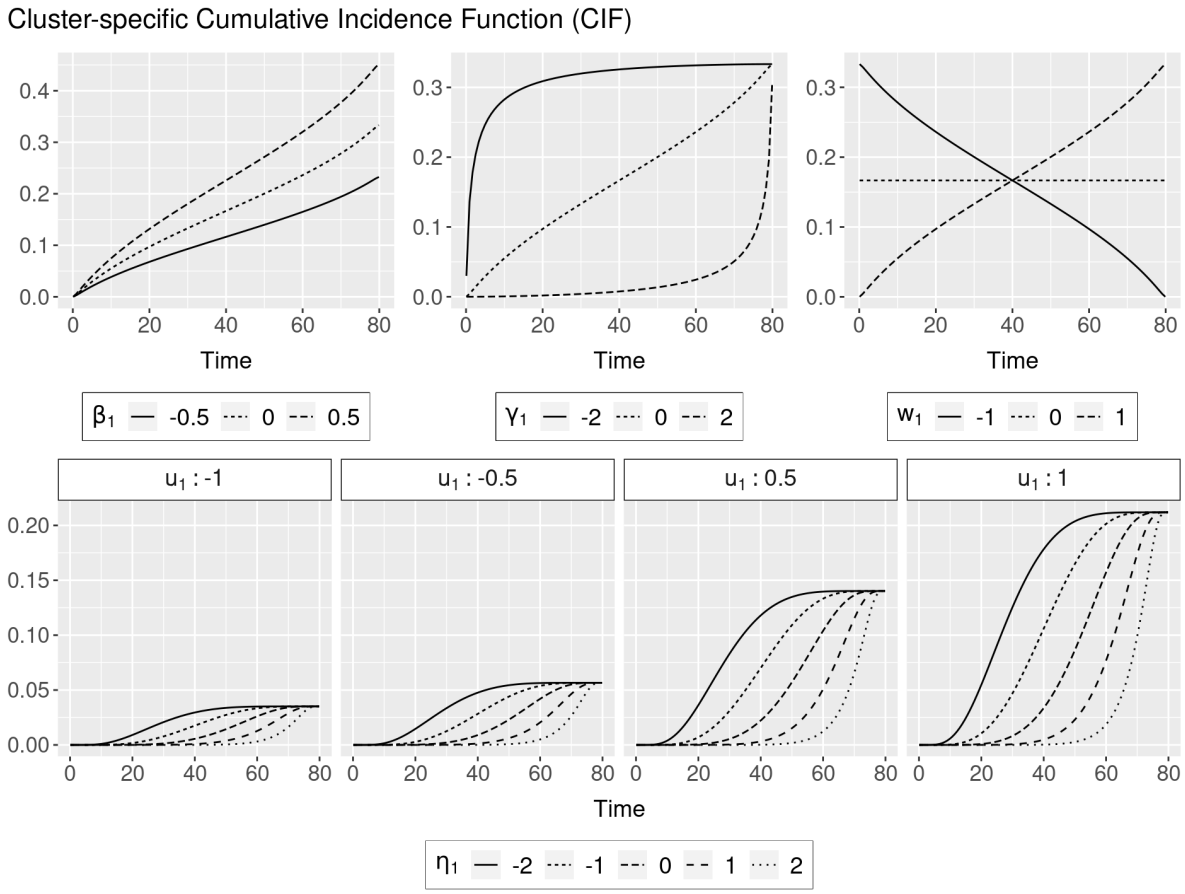

Figure 2: Curve behaviors for different parameter settings, showing then the corresponding parameter effects in a cluster-specific cumulative incidence function (CIF).

Remains to talk about the within-cluster dependence induced by the latent effects in $\boldsymbol{u}$ and $\boldsymbol{\eta}$. To help in the discussion, the bottom-plots on Figure 2 illustrates the cluster-specific CIF for a given failure cause in a model without covariates, let us call it failure cause 1 (in total we have two). The latent effects u_1 and u_2 always appear together in the cluster-specific risk level, as consequency they have a joint effect on the cumulative incidence of both causes. As we can see in Figure 2, an increase in u_k will increase the risk of failure from cause k. The interpretation of $\text{cov}(\eta_1,\ \eta_2)$ and $\text{cov}(u_1,\ u_2)$ is straightforward, and those values are in most of the cases positive, as said in Cederkvist et al. (2019). With regard to $\text{cov}(u_k,\ \eta_k)$, negative values are the common situation. A negative correlation between η_k and u_k imply that when η_k decreases, u_k increases and conversely when η_K increases, u_k decreases. In other words, an increased risk level is

reached quickly and a decreased risk level is reached later, respectively. With regard to cross-cause correlation between η and u, positive values are the common situation where late onset of one failure cause is associated with a high absolute risk of another failure cause.

2.2 Model specification

The generalized linear mixed model (GLMM) for clustered competing risks data is specified in the following hierarchical fashion. By simplicity, we focus on two competing causes of failure but an extension is straightforward. For two competing causes of failure, a subject i, in the cluster/family j and time t, we have

$$y_{ijt} \mid \{u_{1j},\ u_{2j},\ \eta_{1j},\ \eta_{2j}\} \sim \text{Multinomial}(p_{1ijt},\ p_{2ijt},\ p_{3ijt})$$

$$\begin{bmatrix} u_1 \\ u_2 \\ \eta_1 \\ \eta_2 \end{bmatrix} \sim \text{MN}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{u_1} & \text{cov}(u_1,\ u_2) & \text{cov}(u_1,\ \eta_1) & \text{cov}(u_1,\ \eta_2) \\ & \sigma^2_{u_2} & \text{cov}(u_2,\ \eta_1) & \text{cov}(u_2,\ \eta_2) \\ & & \sigma^2_{\eta_1} & \text{cov}(\eta_1,\ \eta_2) \\ & & & \sigma^2_{\eta_2} \end{bmatrix} \right)$$

$$\begin{aligned} p_{kijt} &= \frac{\partial}{\partial t} F_k(t \mid \boldsymbol{x},\ u_1,\ u_2,\ \eta_k) \\ &= \frac{\exp\{\boldsymbol{x}_{kij}\boldsymbol{\beta}_k + u_{kj}\}}{1 + \sum_{m=1}^{K-1} \exp\{\boldsymbol{x}_{mij}\boldsymbol{\beta}_m + u_{mj}\}} \\ &\times w_k \frac{\delta}{2\delta t - 2t^2}\ \phi\left(w_k \text{arctanh}\left(\frac{t - \delta/2}{\delta/2} \right) - \boldsymbol{x}_{kij}\boldsymbol{\gamma}_k - \eta_{kj} \right), \end{aligned} \tag{3}$$

$k = 1,\ 2.$

The probabilities are given by the derivative w.r.t. time t of the cluster-specific CIF. The choice of a multinomial logistic regression model ensures that the sum of all predicted cause-specific CIFs does not exceed 1. Considering two competing causes of failure, we have a multinomial with three classes. The third class exists to handle the censorship and its probability is given by the complementary to reach 1. To better describe these curve behaviors we have Figure 3.

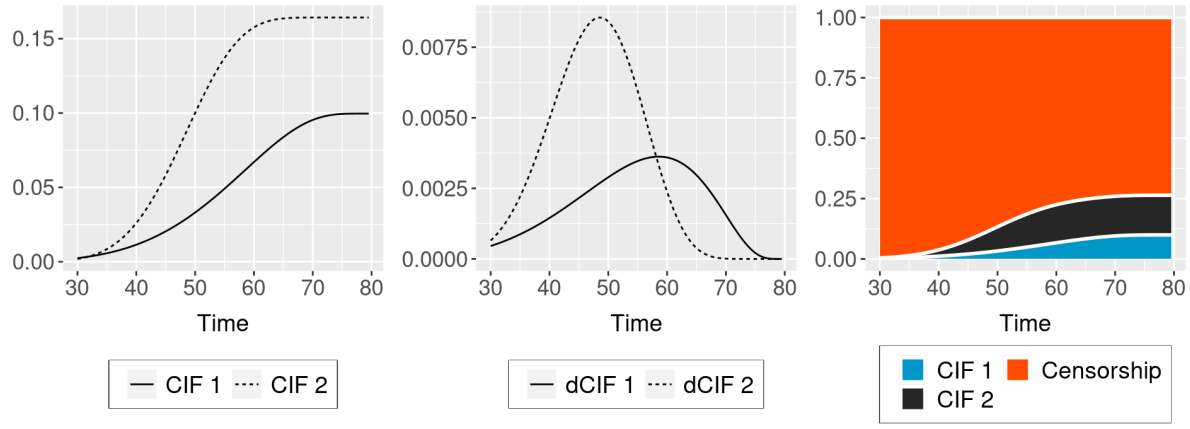

Figure 3: Cluster-specific cumulative incidence function (CIF) curves and their derivatives (dCIF) for a random scenario with two competing causes.

This framework in Equation 3 results in what we call multiGLMM, a multinomial GLMM to handle the CIF of clustered competing risks data. For a random sample, the corresponding marginal likelihood function in given by

$$
\begin{aligned}
L(\boldsymbol{\theta}\ ;\ y) &= \prod_{j=1}^{J} \int_{\Re^4} \pi(y_j \mid \boldsymbol{r}_j) \times \pi(\boldsymbol{r}_j)\ \mathrm{d}\boldsymbol{r}_j \\
&= \prod_{j=1}^{J} \int_{\Re^4} \underbrace{\left\{ \prod_{i=1}^{n_j} \prod_{t=1}^{n_{ij}} \left(\frac{(\sum_{k=1}^{K} y_{kijt})!}{y_{1ijt}!\ y_{2ijt}!\ y_{3ijt}!} \prod_{k=1}^{K} p_{kijt}^{y_{kijt}} \right) \right\}}_{\text{fixed effect component}} \times \\
&\qquad \underbrace{(2\pi)^{-2} |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2} \boldsymbol{r}_j^{\top} \Sigma^{-1} \boldsymbol{r}_j \right\}}_{\text{latent effect component}} \mathrm{d}\boldsymbol{r}_j \\
&= \prod_{j=1}^{J} \int_{\Re^4} \underbrace{\left\{ \prod_{i=1}^{n_j} \prod_{t=1}^{n_{ij}} \prod_{k=1}^{K} p_{kijt}^{y_{kijt}} \right\}}_{\text{fixed effect}} \underbrace{(2\pi)^{-2} |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2} \boldsymbol{r}_j^{\top} \Sigma^{-1} \boldsymbol{r}_j \right\}}_{\text{latent effect component}} \mathrm{d}\boldsymbol{r}_j, \qquad (4)
\end{aligned}
$$

where $\boldsymbol{\theta} = [\boldsymbol{\beta}\ \boldsymbol{\gamma}\ \boldsymbol{w}\ \boldsymbol{\sigma^2}\ \boldsymbol{\rho}]^{\top}$ is the parameters vector to be maximized. In our framework, a subject can fail from just one competing cause or get censor, at a given time. Thus, the fraction of factorials in the fixed effect component is made only by 0's and 1's. The matrix Σ is the variance-covariance matrix, which parameters are given by $\boldsymbol{\sigma}^2$ and $\boldsymbol{\rho}$.

To each cluster j we have a product of two components. The fixed effect component, given by a multinomial distribution with its probabilities specified through the cluster-specific CIF (Equation 1) and, the latent effect component, given by a multivariate Gaussian distribution. To each subject i that composes a cluster j we have its specific fixed effects contribution. The likelihood in Equation 4 is the most general as possible, allowing for repeated measures to each subject. Since all subjects of a given cluster shares the same latent effect, we have just one latent effect contribution multiplying the product of fixed

effect contributions. As we do not observe the latent effect variables, $\boldsymbol{r}_j$, we integrate out in it. With two competing causes of failure we have four latent effects, a multivariate Gaussian distribution in four dimensions. Consequently, for each cluster, we approximate an integral in four dimensions. The product of these approximated integrals results in the called marginal likelihood, to be maximized in $\boldsymbol{\theta}$.

3 Estimation and inference

Our goal is to estimate the parameter vector $\boldsymbol{\theta} = [\boldsymbol{\beta}\ \boldsymbol{\gamma}\ \boldsymbol{w}\ \boldsymbol{\sigma^2}\ \boldsymbol{\rho}]^\top$. The likelihood for $\boldsymbol{\theta}$ can be written as

$$L(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{u}) = \prod_{i=1}^{I} \prod_{j=1}^{n_i} f(y_{ij} \mid \boldsymbol{u}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma})\ f(\boldsymbol{u}_i \mid \boldsymbol{\Sigma}). \tag{5}$$

From standard probability theory is easy to see that in the right-hand side (r.h.s.) we have a joint density, consequently, Equation 5 represents what is called a full or joint likelihood function. The latent effect $\boldsymbol{u}$ is *latent*, i.e. we do not observe it. To handle this we use the Laplace approximation.

If we have a joint density we can just integrate out the undesired variable resulting in

$$\begin{aligned} L(\boldsymbol{\theta} \mid \boldsymbol{y}) &= \prod_{i=1}^{I} \int_{\mathcal{R}^{\boldsymbol{u}_i}} \left[\prod_{j=1}^{n_i} f(y_{ij} \mid \boldsymbol{u}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma})\ f(\boldsymbol{u}_i \mid \boldsymbol{\Sigma}) \right] \mathrm{d}\boldsymbol{u}_i \\ &= \prod_{i=1}^{I} \int_{\mathcal{R}^{\boldsymbol{u}_i}} f(\boldsymbol{y}_i, \boldsymbol{u}_i \mid \boldsymbol{\theta})\ \mathrm{d}\boldsymbol{u}_i, \end{aligned} \tag{6}$$

a marginal density that keeps the parameters $\{\boldsymbol{\sigma^2}\ \boldsymbol{\rho}\}$ of the integrated variable. To handle this integration step a clever choice is to take advantage of the exponential family structure together with the Gaussian latent effects distribution. These ideas converge to an adaptive Gaussian quadrature with one integration point, also known as *Laplace approximation* (Molenberghs and Verbeke; 2005; Shun and McCullagh; 1995; Tierney and Kadane; 1986; Wood; 2015).

We may approximate a analytically intractable integral in a way to obtain a tractable closed-form expression allowing the numerical maximization of the resulting marginal likelihood function (Bonat and Ribeiro Jr; 2016). The Laplace approximation has been designed to approximate integrals in the form

$$\int_{\mathcal{R}^{\boldsymbol{u}_i}} \exp\{Q(\boldsymbol{u}_i)\} \mathrm{d}\boldsymbol{u}_i \approx (2\pi)^{n_{\boldsymbol{u}}/2}\ |Q''(\hat{\boldsymbol{u}}_i)|^{-1/2}\ \exp\{Q(\hat{\boldsymbol{u}}_i)\}, \tag{7}$$

where $Q(\boldsymbol{u}_i)$ is a known unimodal bounded function, and $\hat{\boldsymbol{u}}_i$ is the value for which $Q(\boldsymbol{u}_i)$ is maximized. A advantage of the Laplace approximation approach in a GLMM is the

exponential family structure. In a usual GLMM the response follows a one-parameter exponential family distribution that can be written as

$$f(\boldsymbol{y}_i \mid \boldsymbol{u}_i, \boldsymbol{\theta}) = \exp\left\{\boldsymbol{y}_i^\top(\boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{u}_i) - \mathbf{1}_i^\top b(\boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{u}_i) + \mathbf{1}_i^\top c(\boldsymbol{y}_i)\right\},$$

where $b(\cdot)$ and $c(\cdot)$ are known functions. This general and easy to compute expression together with a (multivariate) Gaussian distribution, highlights the convenience of the Laplace method. The $Q(\boldsymbol{u}_i)$ function to be maximized can be expressed as

$$\begin{aligned} Q(\boldsymbol{u}_i) = {} & \boldsymbol{y}_i^\top(\boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{u}_i) - \mathbf{1}_i^\top b(\boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{u}_i) + \mathbf{1}_i^\top c(\boldsymbol{y}_i) \\ & - \frac{n_{\boldsymbol{u}}}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\boldsymbol{u}_i^\top\boldsymbol{\Sigma}^{-1}\,\boldsymbol{u}_i. \end{aligned} \tag{8}$$

The approximation in Equation 7 requires the maximum $\hat{\boldsymbol{u}}_i$ of the function $Q(\boldsymbol{u}_i)$. As we assume a Gaussian distribution with a known mean for the latent effects, we have the perfect initial guess for a Hessian-based maximization method, as the Newton-Raphson (NR) algorithm. Bonat and Ribeiro Jr (2016) presents the generic expressions for the derivatives required by the NR method, given by the following:

$$\begin{aligned} Q'(\boldsymbol{u}_i^{(k)}) &= \{\boldsymbol{y}_i - b'(\boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{u}_i^{(k)})\}^\top - \boldsymbol{u}_i^{(k)\top}\boldsymbol{\Sigma}^{-1}, \\ Q''(\boldsymbol{u}_i^{(k)}) &= -\text{diag}\{b''(\boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{u}_i^{(k)})\} - \boldsymbol{\Sigma}^{-1}. \end{aligned}$$

The marginal log-likelihood function returned by the Laplace approximation, to each individual or unit under study i, is as follows:

$$\begin{aligned} l(\boldsymbol{\theta} \mid \boldsymbol{y}_i) = \log L(\boldsymbol{\theta} \mid \boldsymbol{y}_i) = {} & \frac{n}{2}\log(2\pi) - \frac{1}{2}\log\left|\text{diag}\{b''(\boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\hat{\boldsymbol{u}}_i)\} + \boldsymbol{\Sigma}^{-1}\right| \\ & + \boldsymbol{y}_i^\top(\boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\hat{\boldsymbol{u}}_i) - \mathbf{1}_i^\top b(\boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\hat{\boldsymbol{u}}_i) + \mathbf{1}_i^\top c(\boldsymbol{y}_i) \\ & - \frac{n_{\boldsymbol{u}}}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\hat{\boldsymbol{u}}_i^\top\boldsymbol{\Sigma}^{-1}\,\hat{\boldsymbol{u}}_i, \end{aligned}$$

that can now be numerically maximized over the model parameters $\boldsymbol{\theta} = [\boldsymbol{\beta}\ \boldsymbol{\gamma}\ \boldsymbol{w}\ \boldsymbol{\sigma^2}\ \boldsymbol{\rho}]^\top$ using a quasi-Newton method as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (Nocedal and Wright; 2006) or the PORT routines (Dennis et al.; 1981; Gay; 1990), all available in the R (R Core Team; 2021) statistical software.

We use an efficient Laplace approximation routine implemented in `TMB` (Kristensen et al.; 2016). Besides state-of-art linear algebra libraries and the possibility of performing the computations in parallel, `TMB` also offers an efficient implementation of automatic differentiation (Nocedal and Wright; 2006; Wood; 2015; Peyré; 2020), the state-of-art in gradients computation. In `TMB` the user should write its likelihood function in a `C++` template file and then load it on R. This last characteristic makes `TMB` general (a vast range of models being allowed to be written) and powerful (low-level `C++` model implementation).

4 Simulation studies

To verify the practical viability of our model we performed an extensive simulation study. As main complicator factors, we may mention the competing causes un balancement, where we typically have much more censorships than actual failures; and the high dimensionality problem, having a considerable number of parameters to estimate in both fixed and latent effects layers.

One of the main questions surrounding our model that we tried to tack down in this study was, if even when in the possession of a high correlated sample in the latent field, we are able to estimate all variance and covariance parameters.

One of the main questions surrounding our model that we tried to tack down in this study was, if even when in the possession of a high correlated sample in the latent field, we are able to estimate all variance and covariance parameters. To answer this question we worked with four different models, with each one differentiating from the other in terms of latent effects structure. We had a model with (i) latent effects only on the risk level; (ii) only on the failure time trajectory level; (iii) on both levels but without cross-correlations, called a blog-diag model; and (iv) a model with all possible correlations presented, called as a complete model. In Figure 4 a visual representation of these latent effects structures is presented, in terms of the corresponding variance-covariance matrices.

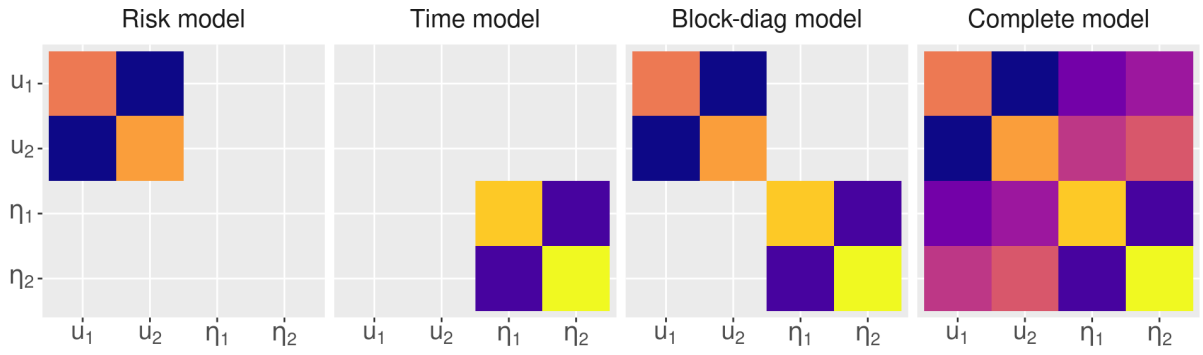

Figure 4: Simulation study variance-covariance matrix model designs, considering two competing causes of failure and consequently, four latent effects.

Besides the four latent effects structures, we worked with two CIF configurations

$$\begin{aligned} \text{High CIF configuration}: &\quad \{\beta_1 = -2,\ \beta_2 = -1.5,\ \gamma_1 = 1,\ \gamma_2 = 1.5,\ w_1 = 3,\ w_2 = 4\}; \\ \text{Low CIF configuration}: &\quad \{\beta_1 = 3,\ \beta_2 = 2.5,\ \gamma_1 = 2.6,\ \gamma_2 = 4,\ w_1 = 5,\ w_2 = 10\}. \end{aligned}$$

Those two sets of parameter values together with the following variance-covariance structure values arise in CIF curves with peaks around 0.15 (low incidence scenario) and 0.60 (high incidence scenario). For numerical efficiency reasons, we worked with the log-variances and the Fisher-z-transformation of the correlation parameters.

$$\begin{aligned}\sigma^2_{u_1} &= 1\\ \sigma^2_{u_2} &= 0.7\\ \sigma^2_{\eta_1} &= 0.6\\ \sigma^2_{\eta_2} &= 0.9,\end{aligned} \qquad \text{Correlation structure } = \begin{array}{c} \begin{array}{cccc} u_1 & u_2 & \eta_1 & \eta_2 \end{array} \\ \begin{pmatrix} 1 & 0.1 & -0.5 & 0.3 \\ & 1 & 0.3 & -0.4 \\ & & 1 & 0.2 \\ & & & 1 \end{pmatrix} \end{array} \begin{array}{c} u_1 \\ u_2 \\ \eta_1 \\ \eta_2 \end{array}.$$

We also considered three cluster sizes (2, 5, and 10) and three sample sizes (5000, 30000, and 60000). We had 72 scenarios (4 latent effects structures $\times$ 2 CIF configurations $\times$ 3 cluster sizes $\times$ 3 sample sizes). For each scenario, we simulate 500 samples. In total, 36000 (72×500) model fittings. In the smallest scenario, we had 500 clusters, in the biggest one, 30000 clusters. The number of clusters is the number of Laplace approximations to be performed. With two competing causes of failure, we have a model with 16 parameters (6 in the fixed effect layer, 3 for each competing cause; and 10 in the latent effects layer, 4 variances, and 6 covariance parameters).

All models were run, in a parallelized fashion, in one of the two following Linux systems:

System 1 12 Intel(R) Core(TM) i7-8750H CPU @2.20GHz processors with 16GB RAM;

System 2 30 Intel(R) Xeon(R) CPU E5-2690v2 @3.00GHz processors and 206GB RAM.

The non-complete models (involving 2D Laplace approximations) are kind of fast, taking always less than 5 minutes of processing. In the most expensive scenarios (30K 4D Laplaces), the complete model takes 30 minutes. In terms of parameters estimation, the non-complete models fail to learn the data. They appear to be not structured enough to capture the data characteristics, showing that a full correlated latent field is really the most appropriate choice.

In the supplementary materials, we have several graphs summarizing the parameters estimate bias. In each figure, we have the estimate bias and its uncertainty described by a Wald-based confidence interval, i.e. $\pm$ 1.96 the bias standard deviation. We chose to use this uncertainty representation based on the point estimates instead of the standard error computations, since in several scenarios the model fails to compute all the standard errors, caused by Hessian numerical instabilities.

When we assume a non-zero cross-correlation structure (complete model), the improvements in terms of bias reduction are huge when compared with the non-complete models. The mean biases get very close to zero, the standard deviations decrease 50% or more, if compared with the non-complete models. and the high CIF scenarios are the ones with a much smaller bias-variance. All this is accomplished through the consideration of the cross-correlations.

In the *simpler* models, with a latent structure just in one level, is hard to see some significant difference between the clusters and sample sizes. With the complete model, in

the other hand, the difference is clear: as we increase the clusters and the sample sizes, the bias-variance decreases. The mean-bias is basically always the same. The biggest bias-variances are obtained in the log-variances. A final remark to be made is about convergences. With the simpler models, not all of them work, having in some scenarios (generally the ones with 60 thousand data points) a 50∼60% convergence rate. With the complete model, basically, almost all fits reach convergence (∼95% performance).

About the implied mean-CIF curves, in Figure 5 we have the high CIF scenario curves and in Figure 6 the low CIF scenario curves. Since for all models we have a latent structure for the within-cluster dependency, the inherent idea is that this also affect the fixed-effect parameter estimates. By taking its average in each of the seventy-two scenarios, we are able to construct the mean CIF curves.

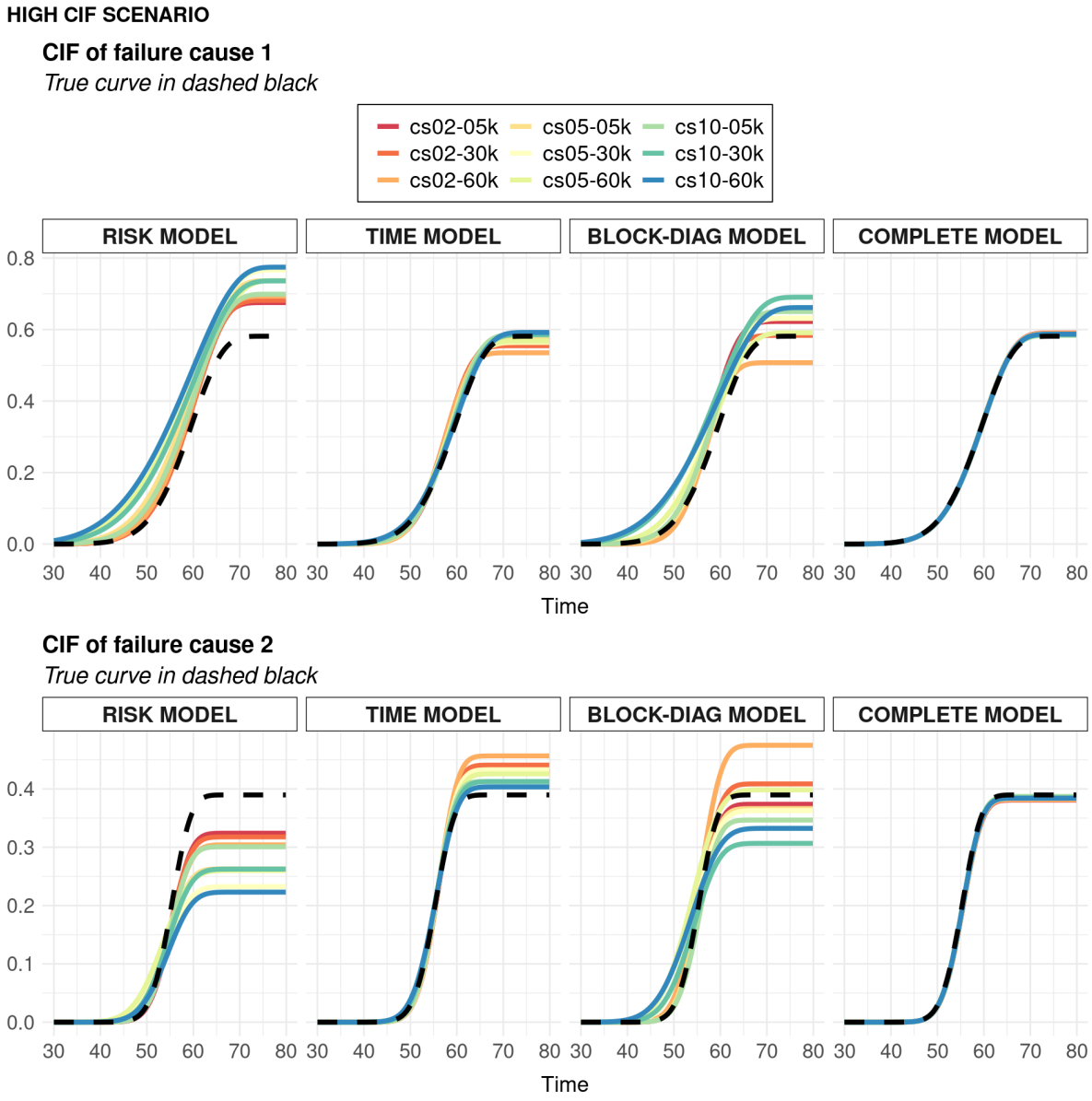

Figure 5: High cumulative incidence function (CIF) scenario curves. `cs` means the cluster size (2, 5, and 10), and 5, 30, and 60K means the sample size.

In Figure 5 is clear that with the complete model we get a perfect fit in all scenarios. The risk and time models estimate well the curve shape parameters but they fail to learn the max incidence. A compensation between curves is clear. In the low CIF scenarios in Figure 6, the estimation is clearly more difficult. The overall fits are bad. For one of the failure causes, the estimation quality is not so bad. The problem is when we look to the other. The best joint fit is still with the complete model.

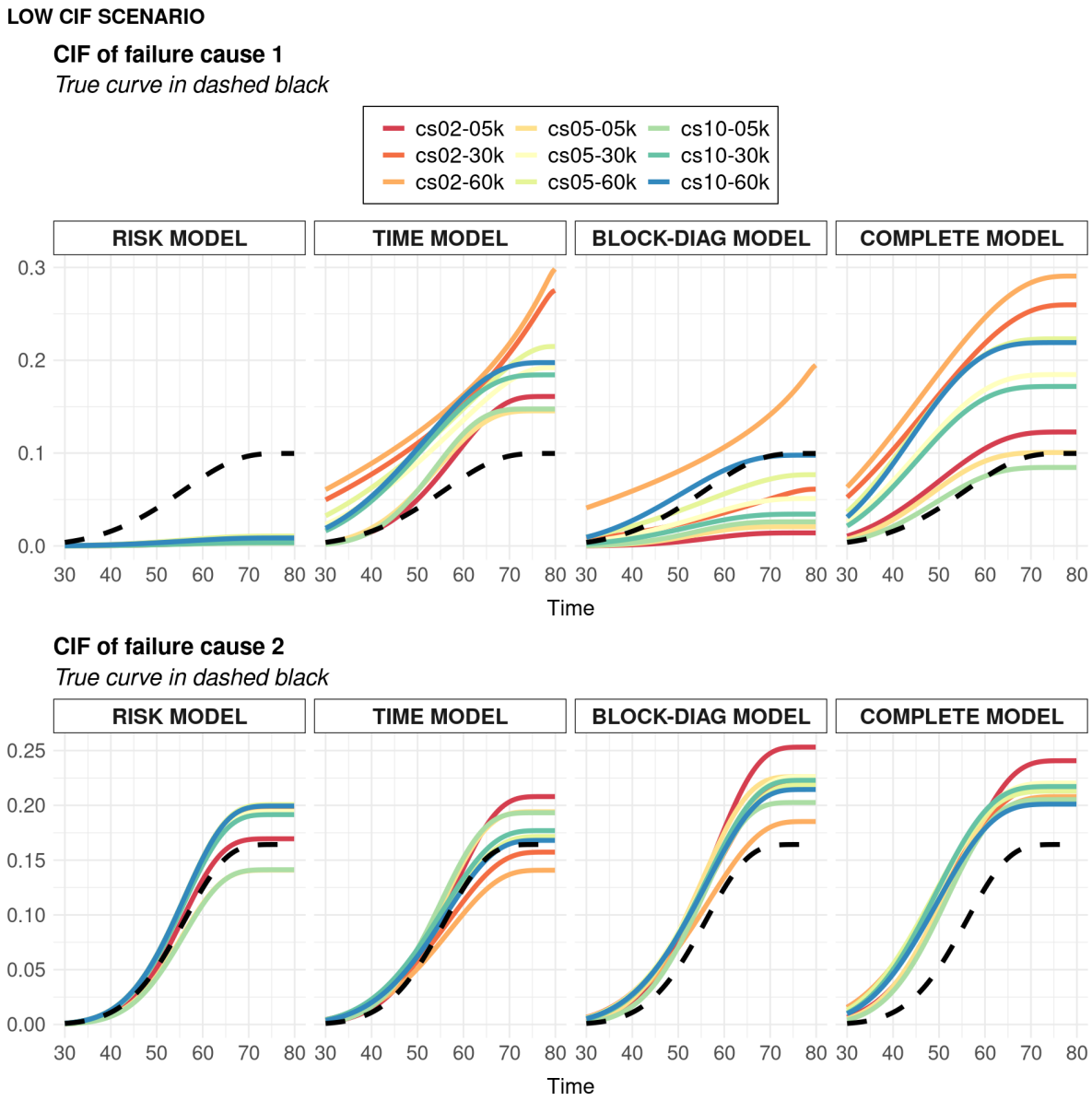

Figure 6: Low cumulative incidence function (CIF) scenario curves. `cs` means the cluster size (2, 5, and 10), and 5, 30, and 60K means the sample size.

5 Discussion

The general goal of this paper was the proposition and evaluation of a maximum likelihood estimation approach for the analysis of clustered competing risks data. Focused on the

probability scale, by means of the cumulative incidence function (CIF), instead of the hazard scale usual in the survival modeling literature (Kalbfleisch and Prentice; 2002). We model the clustered competing risks on a latent-effects framework, a generalized linear mixed model (GLMM) (McCulloch and Searle; 2001), with a multinomial distribution for the competing risks and censorship, conditioned on the latent-effects. The within-cluster latent dependency is accommodated by a multivariate Gaussian distribution and is modeled via its covariance matrix parameters.

The failures by the competing causes and their respective censorships are modeled in the probability scale, by means of the CIF (Kalbfleisch and Prentice; 2002; Andersen et al.; 2012). The CIF is accommodated in our GLMM framework in terms of the link function (McCullagh and Nelder; 1989), as the product of two functions, one responsible to model the instantaneous risk and the other the failure time trajectory, both in a cluster-specific fashion. The shape of these functions is described in detail in Section 2. This particular GLMM formulation is what makes our model, particular. Thus, we have what we call a multiGLMM: a multinomial GLMM for clustered competing risks data.

The two-function product CIF formulation was taken from Cederkvist et al. (2019) but there they use a different estimation framework, a composite likelihood framework (Lindsay; 1988; Cox and Reid; 2004; Varin et al.; 2011). Here we do a full likelihood analysis instead. A composite approach is generally used when a full likelihood approach is impossible or computationally impracticable. Our goal here was to assess a full likelihood framework taking advantage of state-of-the-art computational libraries together with efficient algorithm implementations. We have all this with the R (R Core Team; 2021) package `TMB` (Kristensen et al.; 2016).

The applications in focus here were family studies. Besides the within-cluster/family dependence, this kind of study is characterized by involving big samples, generally, populations. Also, generally having a high number of small clusters, families. A maximum likelihood approach with the use of efficiently implemented Laplace approximations (Tierney and Kadane; 1986; Bonat and Ribeiro Jr; 2016) together with an automatic differentiation (AD) (Wood; 2015; Nocedal and Wright; 2006) routine, all via `TMB`, is able to efficiently handle with a high number of clusters, independent of its size. The multinomial distribution assumption, on its own, is an excellent probabilistic choice since it can accommodate virtually any number of competing causes of failure and its censorship. The presence of those two characteristics in our multiGLMM makes it an efficient and scalable modeling framework for clustered competing risks data.

Even with our modeling framework being virtually able to handle any number of competing causes of failure, we restrained ourselves to work here with only two of them. With two competing causes, we have a 4×4 covariance matrix for the latent effects, which implies ten covariance parameters, which is already a lot of parameters to be estimated in a latent structure. Since our goal was to assess the viability of the maximum likelihood

estimation method, we kept it with two causes.

Finally, the complete model. In the biggest scenario, with 60 thousand data points and clusters of size 2 i.e., with 30 thousand four-dimension integral approximations (ten parameters in the covariance matrix), the model fitting takes 30 minutes, in parallel, with TMB. Before doing the TMB implementation, to really understand what we were doing, we did a complete `R` implementation. We wrote the marginal log-likelihood in `R`, based on our own Laplace approximation Bonat and Ribeiro Jr (2016) and Newton-Raphson implementation (the gradients, **??**, and Hessian, **??**, were computed by hand and implemented). Running this complete `R` implementation in a scenario with 20 thousand data points and clusters of size 2, took around 30 hours, parallelizing it between all threads of system 1. In summary, by using TMB we were able to increase the model size 3 times and to decrease the computational time 60 times. An incredible performance gain.

In a full R implementation with 10K 4D Laplaces, it took 30hrs. TMB is fast.

We also did a Bayesian analysis via Stan/NUTS-HMC (Stan Development Team 2020). 1 week of parallelized processing for a 2500 size 2 clusters scenario with tuned NUTS. This just reinforces the MCMC impracticability for some complex models.

Still, with the complete model, we performed a Bayesian analysis via `tmbstan` **?**. `tmbstan` enables MCMC sampling **??** from a TMB model object using Stan **??**. Sampling can be performed with or without a Laplace approximation for the random effects, based on the probably state-of-art MCMC sampler algorithm, a Hamiltonian Monte Carlo (HMC) algorithm with the No-U-Turn Sampler (NUTS) extension **?**. We performed just one Bayesian model fitting in a modest scenario with 5 thousand data points and clusters of size 2. It took around 1 whole week of parallelized processing in system 1. The results were basically the same as the ones obtained with TMB but this high computational time just reinforces the, still, MCMC framework limitation.

An important point to be made here is about TMB's memory consumption. As the sample size increases, the dimension of the model matrices also increases. This, summed to a high number of clusters (Laplace approximations to be performed), turns out to be a computational nightmare. For several models, even the 16GB RAM of system 1 was not enough. The bottleneck appears to be in the AD tape, which is made in parallel, by default, if the model fitting is in parallel. By turning this option off (line 11 of **??** (**??**) code), we were able to save a lot of memory, making several models practicable.

Model the CIF of clustered competing risks data is far from being trivial or straightforward. The formulation in Equation 1 implies the desired curve behavior, **??**. However, in counterpart, its derivatives w.r.t. time, generates very small probabilities for the failure competing causes, ending by concentrating almost everything on censorship, **??**. For each competing cause with poor data representativity, we have three curve shape parameters to estimate, implying the necessity of having a lot of data to then have enough information about the causes.

We proposed for our multiGLMM an ideally complete latent-effects formulation i.e., correlated latent effects on both levels, instantaneous risk and failure time trajectory. The main underlying idea of the **??** simulation study was to see in which scenarios we would be able to learn all the involved mean and covariance parameters. As part of that, simpler formulations were proposed i.e., latent-effects in only one level, or in both but without cross-correlations. As result, we got that latent effects only in the risk level did not work. The optimization appears to get lost as if something is missing. Inserting latent effects only in the failure time trajectory level returned better results, but still not satisfactorily good. In most of the evaluated scenarios, the block-diagonal model appeared to be in the middle of them, as a compromise. The best results (smallest parameter estimates biases) were obtained with the complete model i.e. when we consider the cross-correlations between levels. In general, we still observe some high variances between the parameter estimates, but given all the problem characteristics mentioned earlier, sounds to be reasonable. On average, the complete model works fine, mainly in the scenarios of high CIF configuration, and also as expected, as the sample size increases. We can also say that as the cluster size increases, the estimates get better but we did not have very strong results supporting that.

6 ADDITIONAL CONSIDERATIONS

The next step was to compare our results with the ones obtained in SCHEIKE, with the composite approach. In the GitHub repository https://github.com/kkholst/mcif/ the authors provide their code. In `mcif/inst/examples/datasim.R` they show how to simulate from the model, and in `mcif/src/loglik.cpp` they have their marginal log-likelihood function. We tried to optimize their marginal log-likelihood over its parameters using basically all `R` `base::optim()` and `base::nlminb()` available methods, in the paper was used the BFGS, one of them. We made several scenarios, using their own simulation scripts and ours, and to our surprise, the model basically does not work.

The optimization in its majority fails, via any gradient-based algorithm (BFGS Nocedal and Wright (2006), PORT Gay (1990); Dennis et al. (1981), conjugate gradient (CG) **?**), generally by Hessian matrix instability problems, a problem which our model also suffers from when we try to compute the parameter estimates standard errors. When the model works, it is because we are using the parameter true values as initial guesses i.e. if the algorithm needs to walk on the log-likelihood surface following the gradient, it fails. Even when it works, the estimates are not always good. We also tried with a SANN and a Nelder-Mead algorithm. SANN **?** is a variant of a simulated annealing method, based on a Metropolis algorithm. Since it is based on simulation, it takes a lot of time and as the gradient-based methods, do not work most of the time. The best results were with the Nelder-Mead **?**, a gradient-free method. Still, it only works when we use the parameter

true values as initial guesses. This situation is completely the opposite of what is shown in the paper, making impossible any reasonable comparison between the models. We will enter in contact with the authors to see what is happening.

7 FUTURE WORKS

As show in **??** results, even with the complete model specification, the parameter estimates present an excessive variance. In terms of a traditional GLMM specification McCulloch and Searle (2001), we do not have a lot more to do. We are already using a smart quasi-Newton algorithm Dennis et al. (1981), the most efficient derivatives computation technique (AD) Peyré (2020), and an also efficient Laplace approximation routine Wood (2015); Bonat and Ribeiro Jr (2016), via TMB Kristensen et al. (2016). We could change the Laplace approximation for an adaptative Gaussian quadrature **?**, but we do not see any good reason to do that.

There are two possible paths here. We could instead of a conditional modeling framework (GLMM/latent-effects model), employ a marginal modeling framework. In this framework, instead of caring about the specification of a probability distribution to the competing causes conditioned on the latent effects, we just care about the specification of a mean and a variance structure. This approach does not have a likelihood function per se, but the estimation procedure tends to be easier than with the GLMM one. A marginal modeling framework that can be used here is the multivariate covariance generalized linear model (McGLM) **??**. How to exactly model the CIF of clustered competing risks data in this framework, is something to still be figured out.

The other path is by the use of a different way of modeling the dependence structure. Instead of a latent-effects approach, we could use copulas **?**Scheike et al. (2010); **?**); **?**. How to do that is something to still be figured out by us, in terms of which kind (conditional or marginal) and version (Archimedean-, Gauss-, Maltesian-, *t*-, hyperbolic-, zebra-, and elliptical-) of copula to use, besides the estimability issue.

Supplementary material

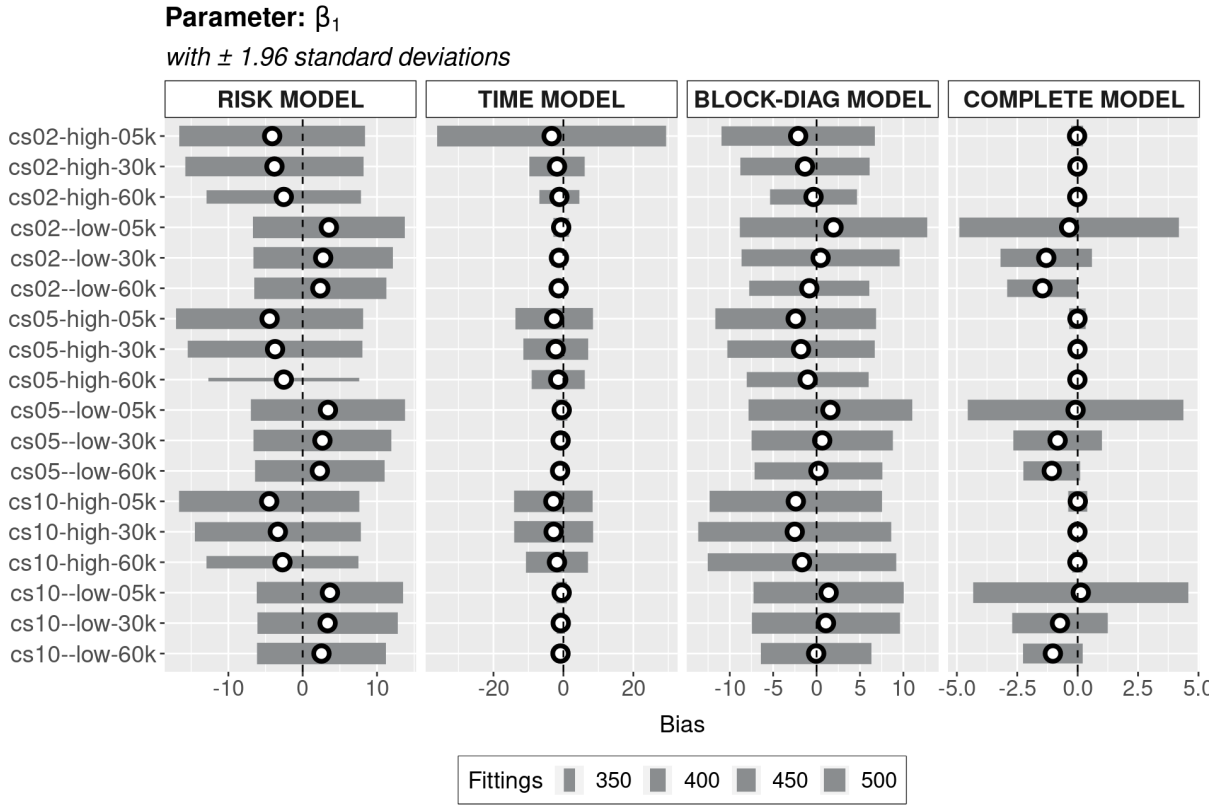

Figure 7: Parameter β_1 bias with $\pm$ 1.96 standard deviations.

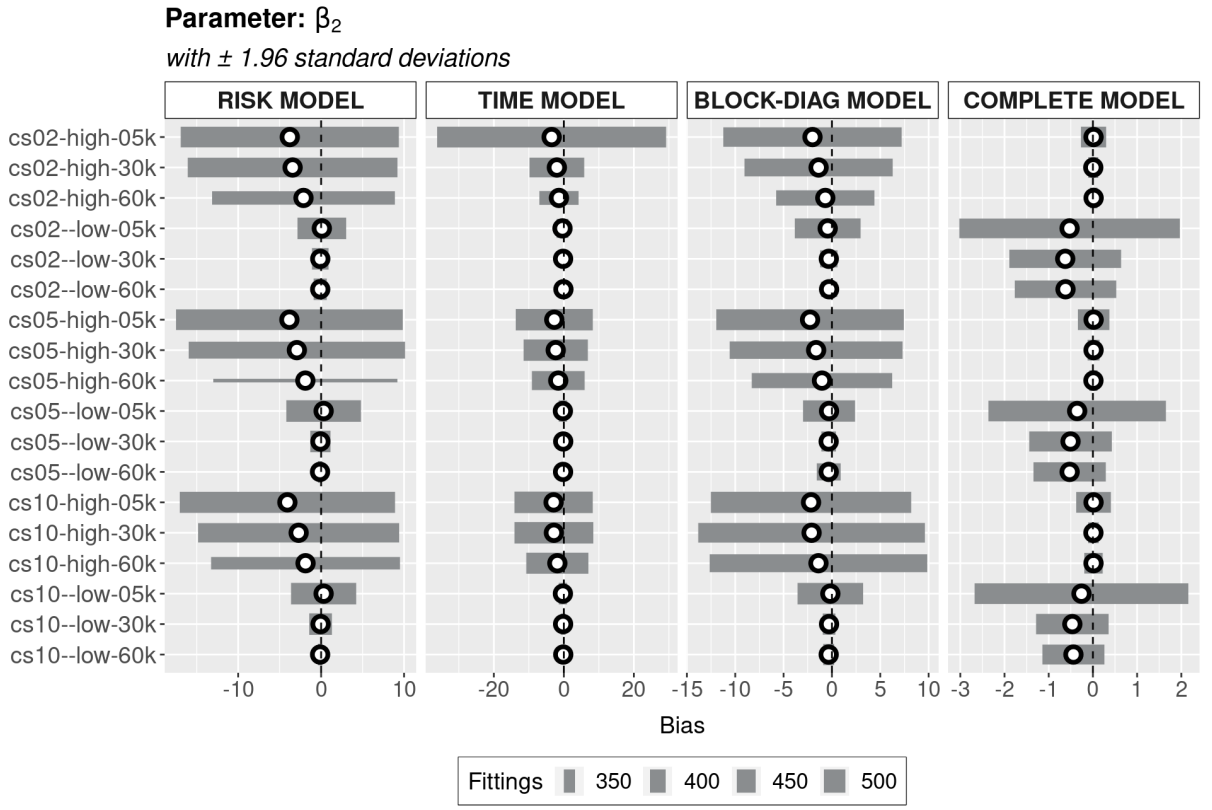

Figure 8: Parameter β_2 bias with $\pm$ 1.96 standard deviations.

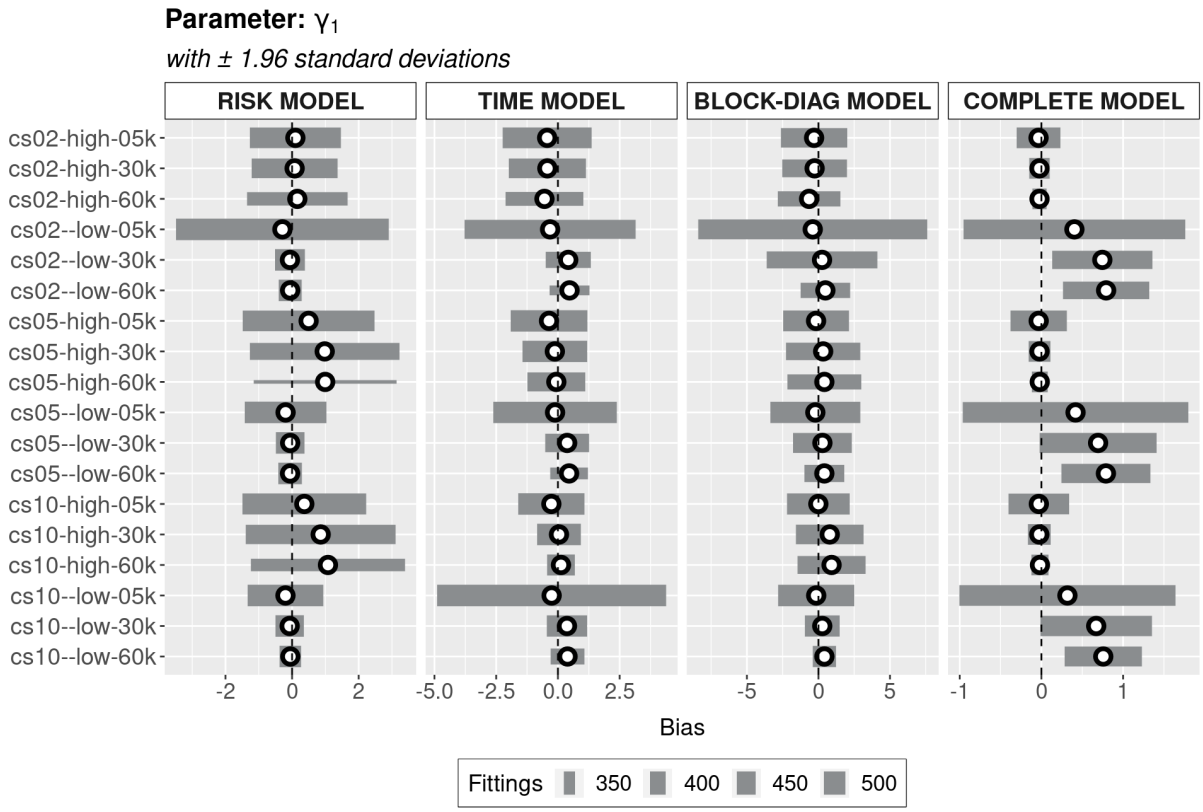

Figure 9: Parameter γ_1 bias with $\pm$ 1.96 standard deviations.

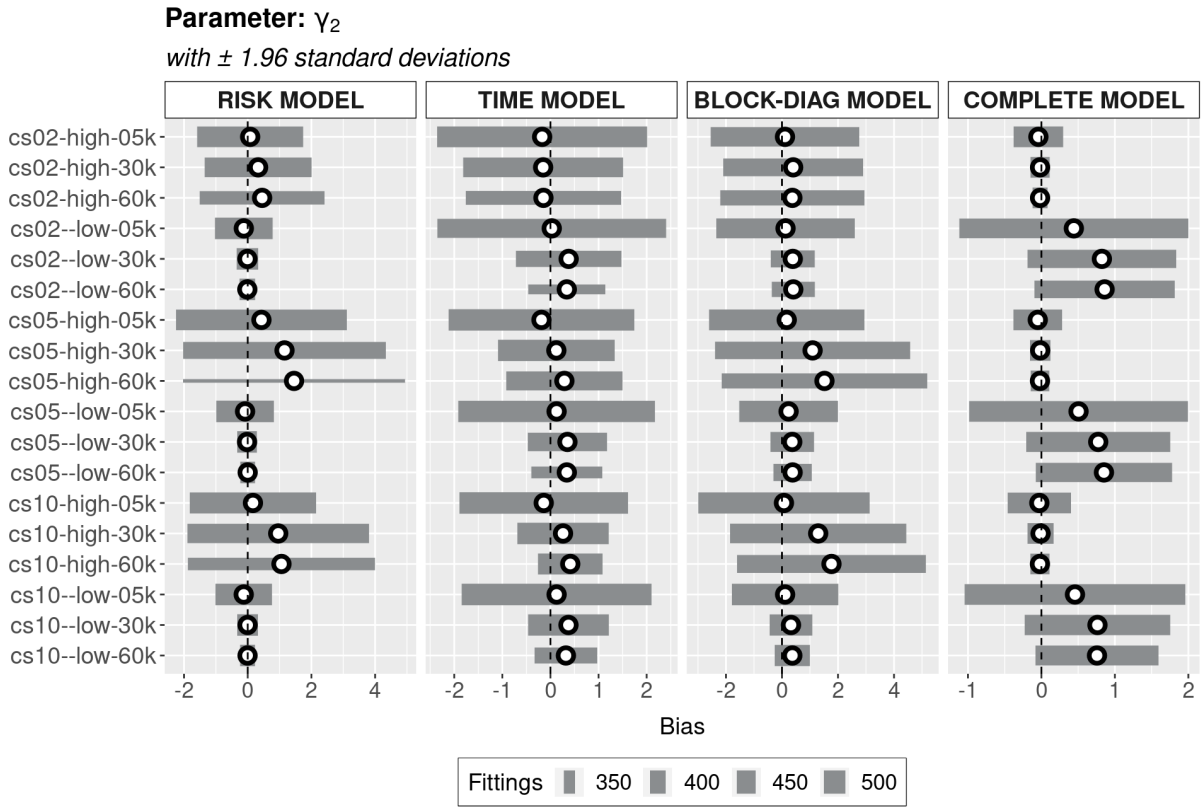

Figure 10: Parameter γ_2 bias with $\pm$ 1.96 standard deviations.

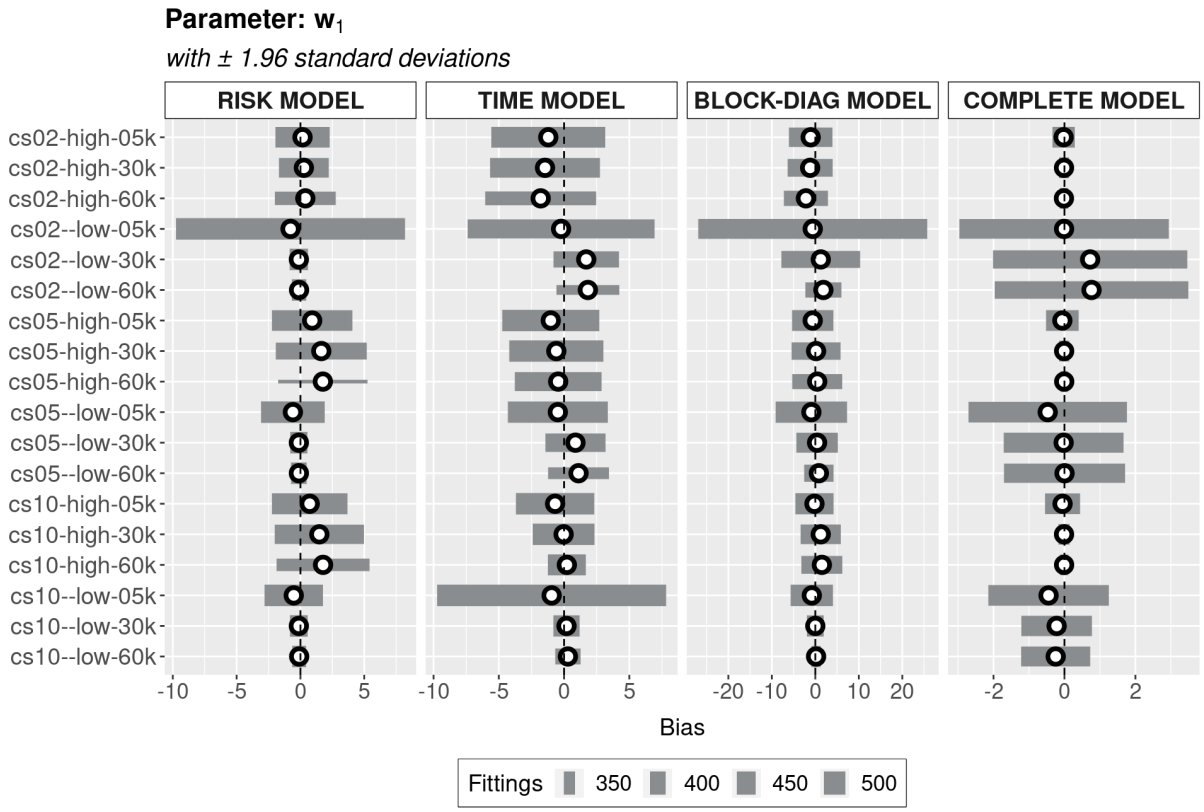

Figure 11: Parameter w_1 bias with ± 1.96 standard deviations.

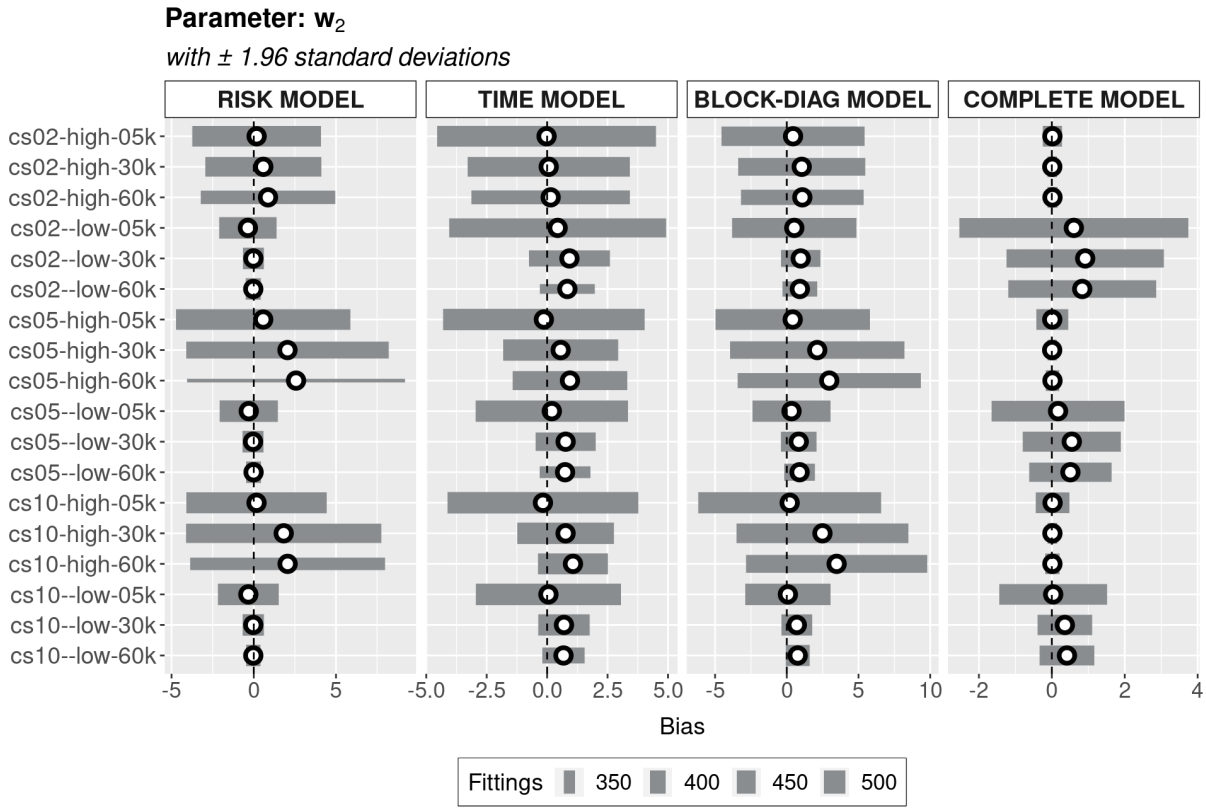

Figure 12: Parameter w_2 bias with ± 1.96 standard deviations.

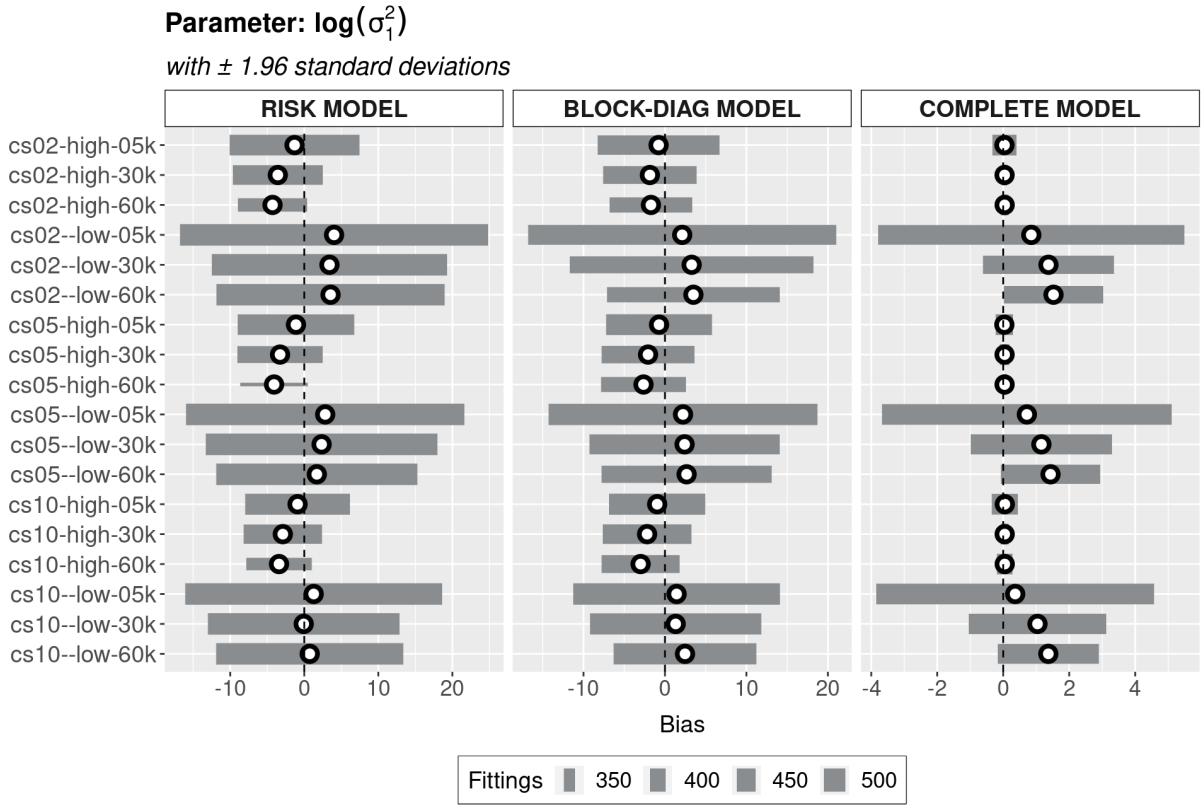

Figure 13: Parameter $\log(\sigma_1^2)$ bias with $\pm$ 1.96 standard deviations.

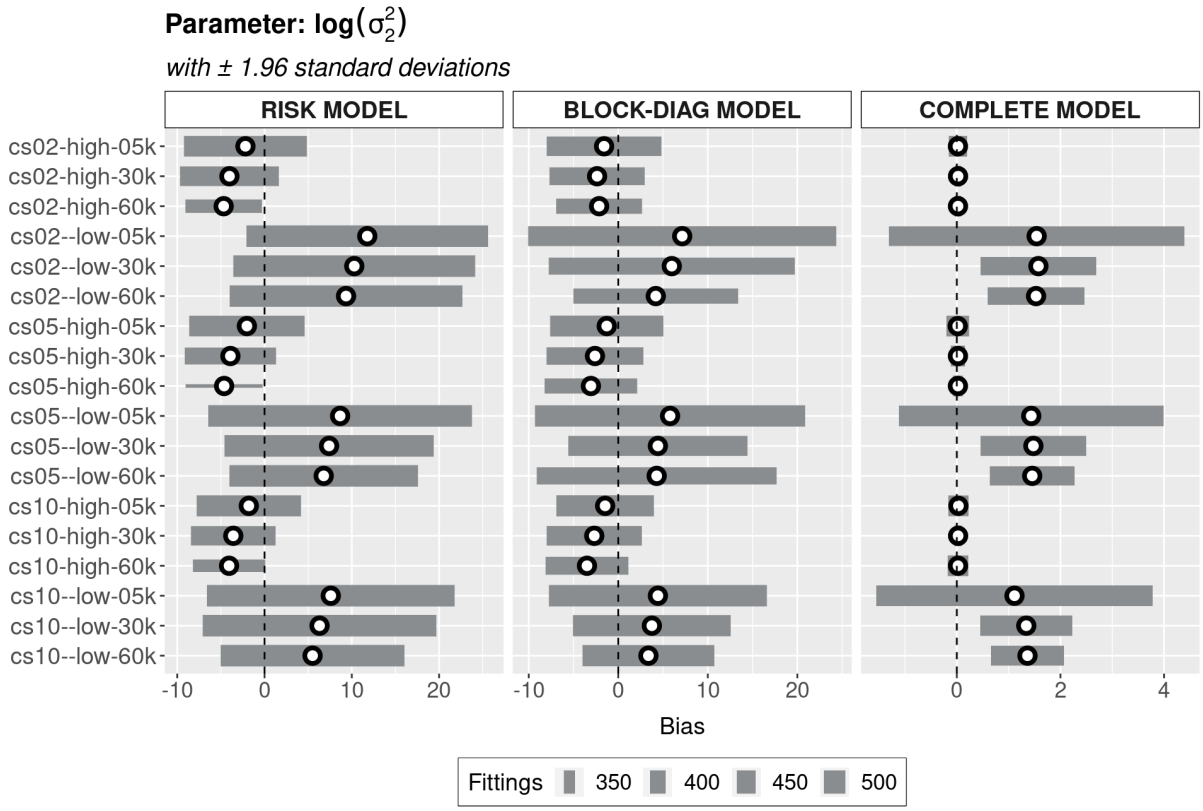

Figure 14: Parameter $\log(\sigma_2^2)$ bias with $\pm$ 1.96 standard deviations.

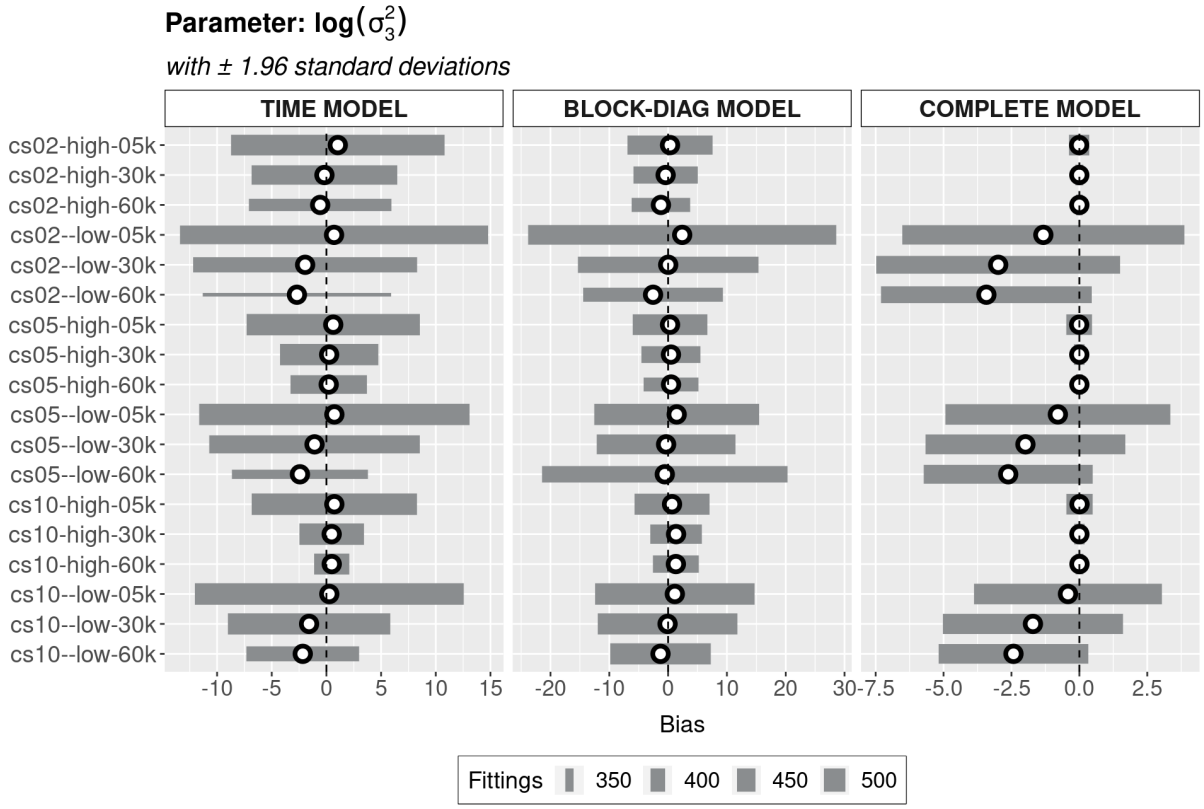

Figure 15: Parameter $\log(\sigma_3^2)$ bias with $\pm$ 1.96 standard deviations.

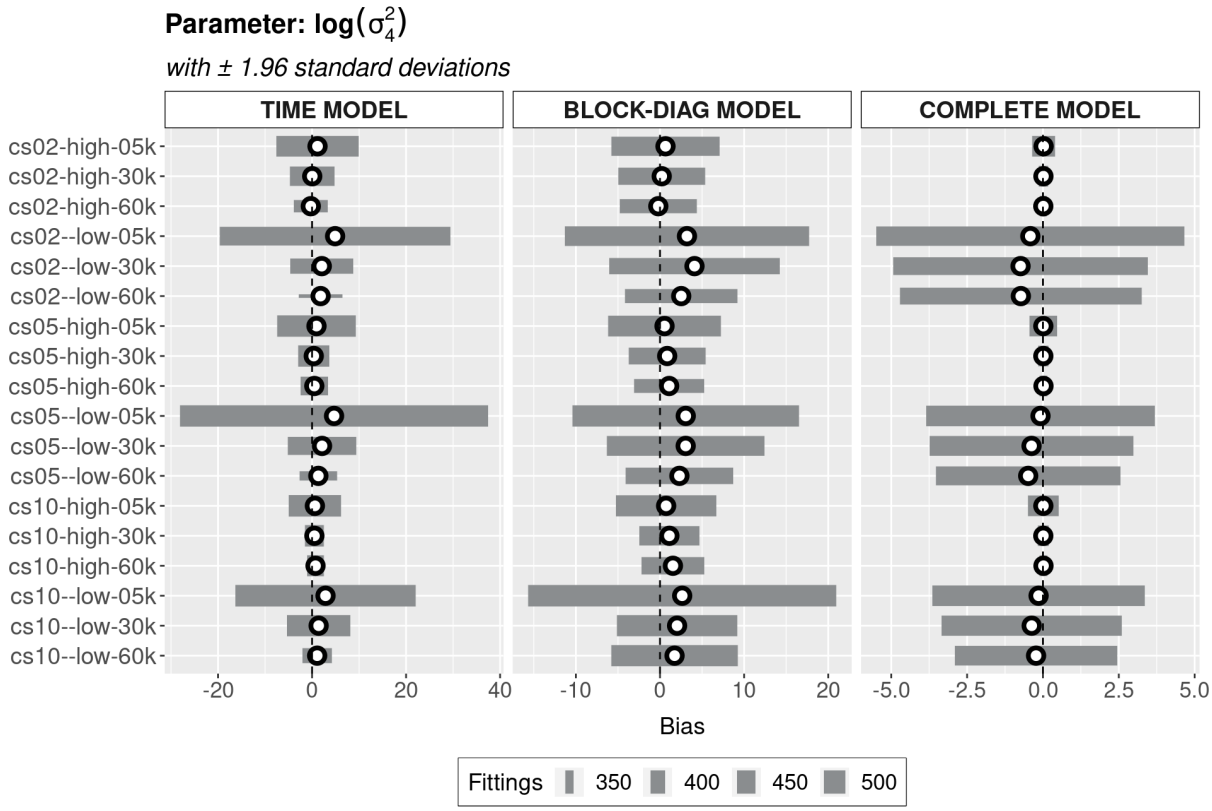

Figure 16: Parameter $\log(\sigma_4^2)$ bias with $\pm$ 1.96 standard deviations.

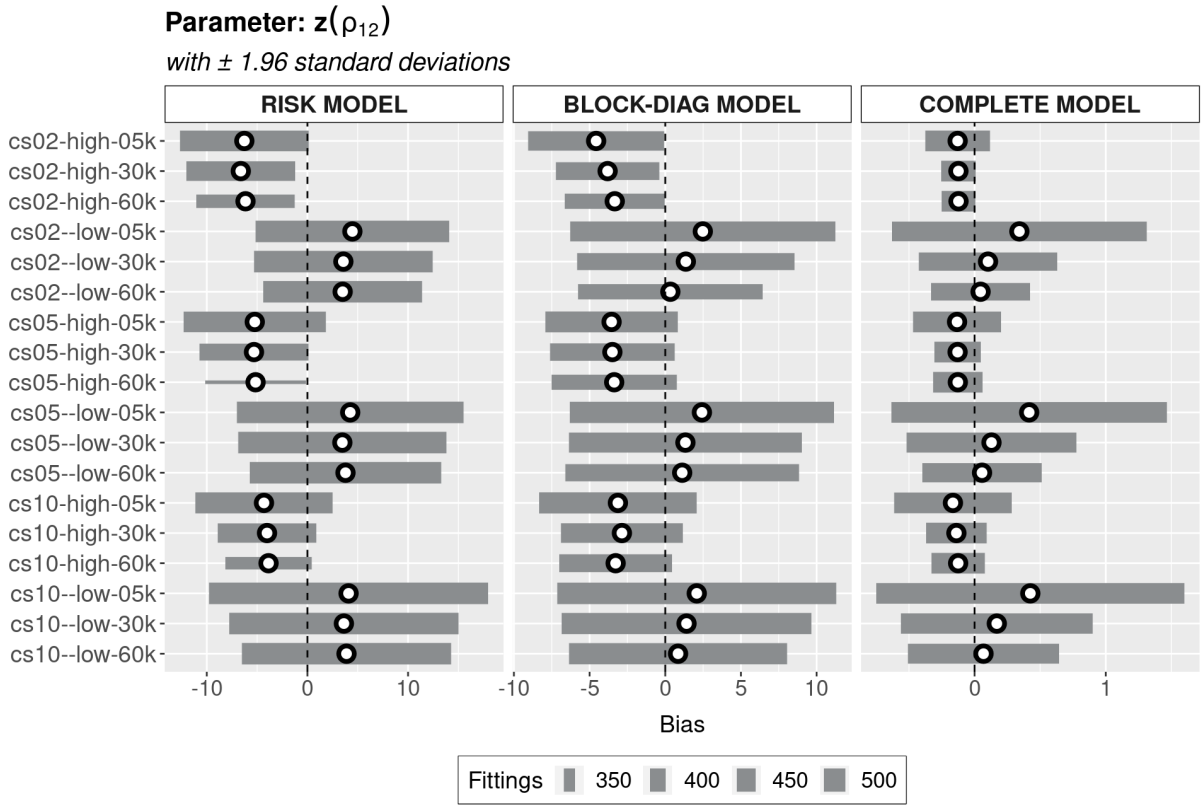

Figure 17: Parameter $z(\rho_{12})$ bias with $\pm$ 1.96 standard deviations.

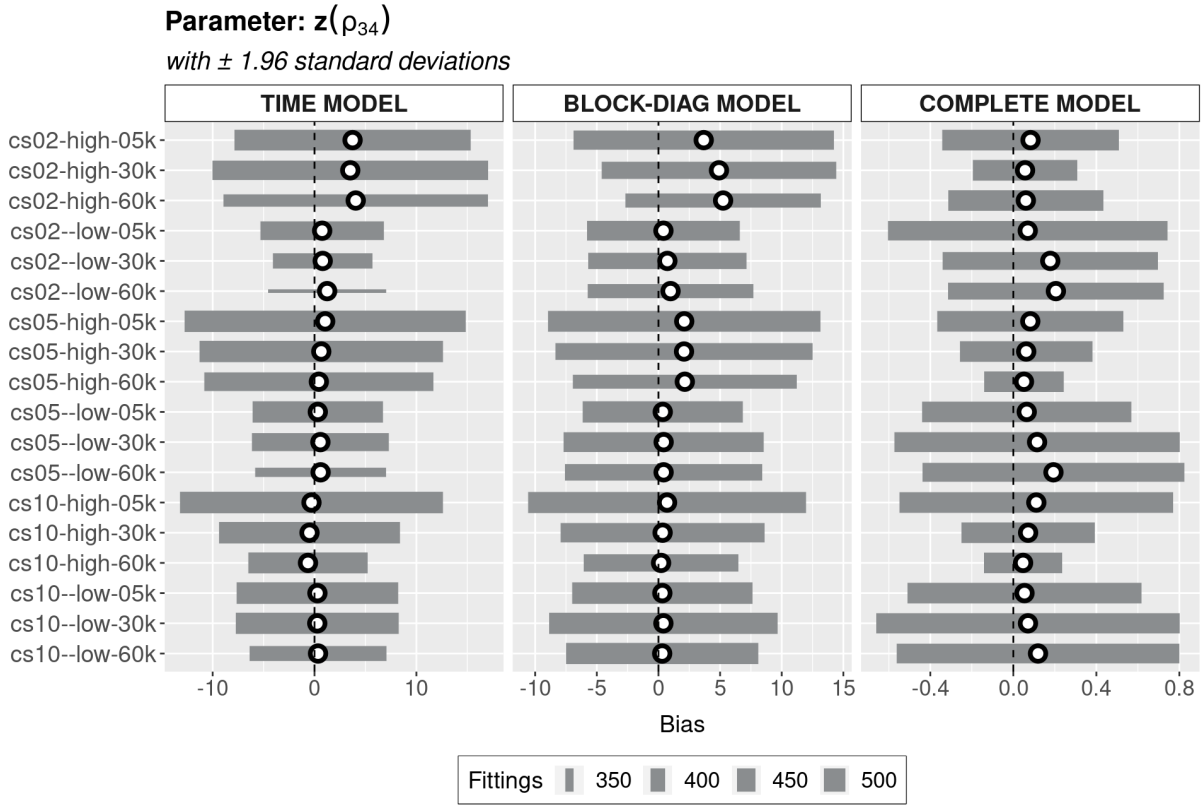

Figure 18: Parameter $z(\rho_{34})$ bias with $\pm$ 1.96 standard deviations.

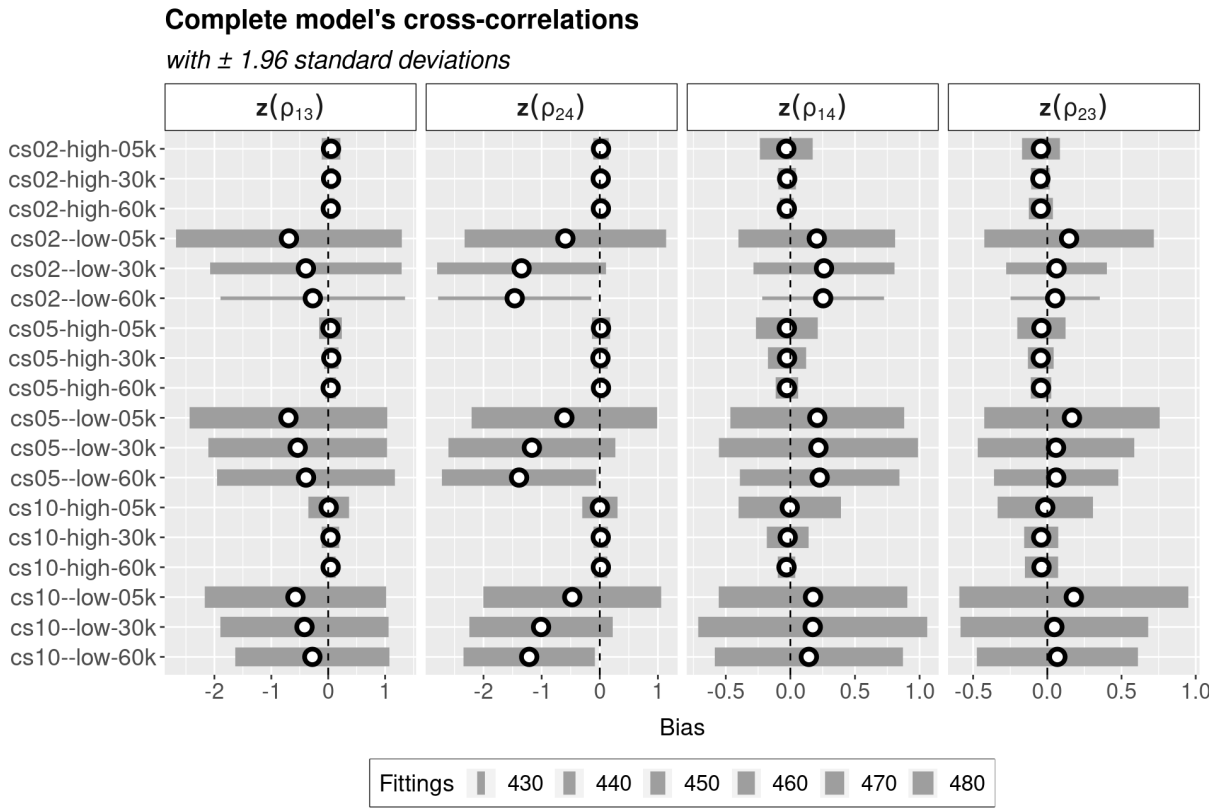

Figure 19: Parameter $\{z(\rho_{13}),\ z(\rho_{24}),\ z(\rho_{14}),\ z(\rho_{23})\}$ bias with $\pm$ 1.96 standard deviations.

References

Andersen, P. K., Geskus, R. B., de Witte, T. and Putter, H. (2012). Competing risks in epidemiology: possibilities and pitfalls, *International Journal of Epidemiology* **31**(1): 861–870.

Bonat, W. H. and Ribeiro Jr, P. J. (2016). Practical likelihood analysis for spatial generalized linear mixed models, *Environmetrics* **27**(1): 83–89.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association* **88**(421): 9–25.

Cederkvist, L., Holst, K. K., Andersen, K. K. and Scheike, T. H. (2019). Modeling the cumulative incidence function of multivariate competing risks data allowing for within-cluster dependence of risk and timing, *Biostatistics* **20**(2): 199–217.

Cheng, Y. and Fine, J. P. (2012). Cumulative incidence association models for bivariate competing risks data, *Journal of the Royal Statistical Society, Series B (Methodological)* **74**(2): 183–202.

Cheng, Y., Fine, J. P. and Kosorok, M. R. J. (2007). Nonparametric Association Analysis

of Bivariate Competing-Risks Data, *Journal of the American Statistical Association* **102**(480): 1407–1415.

Cheng, Y., Fine, J. P. and Kosorok, M. R. J. (2009). Nonparametric Association Analysis of Exchangeable Clustered Competing Risks Data, *Biometrics* **65**(1): 385–393.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial rendency in chronic disease incidence, *Biometrika* **65**(1): 141–151.

Cox, D. R. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities, *Biometrika* **91**(3): 729–737.

Dennis, J. E., Gay, D. M. and Welsch, R. E. (1981). An Adaptive Nonlinear Least-Squares Algorithm, *ACM Transactions on Mathematical Software* **7**(3): 348–368.

Fine, J. P. (1999). Analysing competing risks data with transformation models, *Journal of the Royal Statistical Society, Series B (Methodological)* **61**(4): 817–830.

Fine, J. P. and Gray, R. J. (1999). A proportional hazards models for the subdistribution of a competing risk, *Journal of the American Statistical Association* **94**(446): 496–509.

Gay, D. M. (1990). Usage summary for selected optimization routines, *Technical report*, Computing Science Technical Report 153, AT&T Bell Laboratories. Murray Hill, NJ.

Gerds, T. A., Scheike, T. H. and Andersen, P. K. (2012). Absolute risk regression for competing risks: interpretation, link functions and prediction, *Statistics in Medicine* **31**(29): 3921–3930.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*, Springer-Verlag, New York.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, second Edition edn, John Wiley & Sons, Inc., Hoboken, New Jersey.

Klein, J. P. (1992). Semiparametric estimation of random effects using cox model based on the em algorithm, *Biometrics* **48**(1): 795–806.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H. J. and Bell, B. M. (2016). TMB: Automatic Differentiation and Laplace Approximation, *Journal of Statistical Software* **70**(5): 1–21.

Kuk, A. Y. C. (1992). A semiparametric mixture model for the analysis of competing risks data, *Australian Journal of Statistics* **34**(2): 169–180.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data, *Biometrics* **38**(4): 963–974.

Larson, M. G. and Dinse, G. E. (1985). A Mixture Model for the Regression Analysis of Competing Risks Data, *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **34**(3): 201–211.

Liang, K. Y., Self, S., Bandeen-Roche, K. J. and Zeger, S. L. (1995). Some recent developments for regression analysis of multivariate failure time data, *Lifetime Data Analysis* **1**(1): 403–415.

Lindsay, B. G. (1988). Composite likelihood methods, *Comtemporary Mathematics* **80**(1): 221–239.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, second edition edn, Chapman & Hall, London.

McCulloch, C. E. and Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*, John Wiley & Sons, Inc., New York.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*, Springer, New York.

Naskar, M., Das, K. and Ibrahim, J. G. (2005). A Semiparametric Mixture Model for Analyzing Clustered Competing Risks Data, *Biometrics* **61**(3): 729–737.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models, *Journal of the Royal Statistical Society, Series A* **135**(3): 370–384.

Nielsen, G. G., Gill, R. D., Andersen, P. K. and Sørensen, T. I. A. (1992). A Counting Process Approach to Maximum Likelihood Estimation in Frailty Models, *Scandinavian Journal of Statistics* **19**(1): 25–43.

Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*, Springer Series in Operations Research and Financial Engineering, second Edition edn, Springer, New York.

Petersen, J. H. (1998). An Additive Frailty Model for Correlated Life Times, *Biometrics* **54**(1): 646–661.

Peyré, G. (2020). Course notes on optimization for machine learning, May 10, https://mathematical-tours.github.io/book-sources/optim-ml/OptimML.pdf. CNRS & DMA, École Normale Supérieure.

Prentice, R. L., Kalbfleisch, J. D., Peterson Jr, A. V., Flournoy, N., Farewell, V. T. and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks, *Biometrics* **1**(1): 541–554.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Scheike, T. and Sun, Y. (2012). On cross-odds ratio for multivariate competing risks data, *Biostatistics* **13**(4): 680–694.

Scheike, T., Zhang, Y. S. M. and Jensen, T. K. (2010). A semiparametric random effects model for multivariate competing risks, *Biometrika* **97**(1): 133–145.

Shi, H., Cheng, Y. and Jeong, J. H. (2013). Constrained parametric model for simultaneous inference of two cumulative incidence functions, *Biometrical Journal* **55**(1): 82–96.

Shih, J. H. and Albert, P. S. (2009). Modeling Familial Association of Ages at Onset of Disease in the Presence of Competing Risk, *Biometrics* **66**(4): 1012–1023.

Shun, Z. and McCullagh, P. (1995). Laplace approximation of high dimentional integrals, *Journal of the Royal Statistical Society, Series B (Methodological)* **57**(4): 749–760.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*, Springer-Verlag, New York.

Tierney, L. and Kadane, J. (1986). Accurate approximations for posterior moments and marginal densities, *Journal of the American Statistical Association* **81**(393): 82–86.

Valpel, J. W., Manton, K. G. and Stallard, E. (1979). The impact of heterogeneity in Individual Frailty on the Dynamics of Mortality, *Demography* **16**(1): 439–454.

Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods, *Statistica Sinica* **21**(1): 5–42.

Wood, S. N. (2015). *Core Statistics*, Institute of Mathematical Statistics, Textbooks, IMS.