Modeling the cumulative incidence function of clustered competing risks data: a multinomial GLMM approach

Semana Acadêmica de Estatística, UFPR

Henrique Laureano (.github.io)
IPPPP & LEG @ UFPR

September 13, 2021

Giving context: defining where we are and what we did

Object

- Handle clustered competing risks data (a kind of failure time data) through the cumulative incidence function (CIF).

Goal

- Perform maximum likelihood estimation in terms of a full likelihood formulation based on Cederkvist et al. (2019)'s CIF specification (**Scheike's**).

Contribution

- The full likelihood formulation is in terms of a generalized linear mixed model (GLMM) - a conditional approach (with fixed and random/latent effects);
- The optimization and inference are tacked down via an efficient model implementation with the use of *state-of-art* computational libraries (Kristensen et al. (2016)'s TMB).

Clustered competing risk data

Key ideas:

1. Clustered: groups with a dependence structure (e.g. families);
2. Causes competing by *something*;
3. Occurrence time of this *something*.

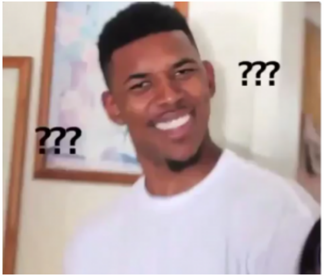
Something?

- **Failure** of an industrial or electronic component;
- **Occurrence** or **cure** of a disease or some biological process;
- **Progress** of a patient clinic state.

Independent of the application, always the same framework

Cluster	ID	Cause 1	Cause 2	Censorship	Time	Feature
1	1	Yes	No	No	10	A
1	2	No	No	Yes	8	A
2	1	No	No	Yes	7	B
2	2	No	Yes	No	5	A

Modeling clustered competing risks data

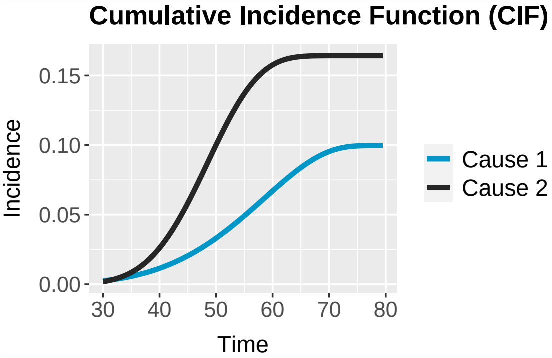


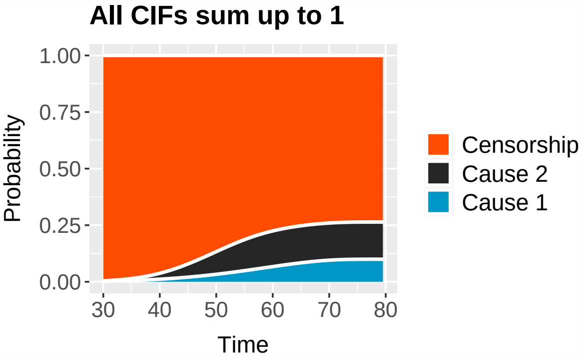
What?

Why?

How?

Probability scale $\rightarrow$ Cause-specific CIF





i.e., CIF $= \mathbb{P}[$ failure time $\leqslant t,$ a given cause $|$ features & latent effects $]$.

Common applications: *family studies*.

↳ Keywords: *within-family/cluster dependence*; *age at disease onset*; *populations*.

Cederkvist et al. (2019)'s CIF specification

For two competing causes of failure,
the cause-specific CIFs are specified in the following manner

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

with

1. $\pi_k(\boldsymbol{x}, \boldsymbol{u}) = \exp\{\boldsymbol{x}\beta_k + u_k\} \Big/ \left(1 + \sum_{m=1}^{K-1} \exp\{\boldsymbol{x}\beta_m + u_m\}\right), \quad k = 1,\ 2, \quad K = 3;$
2. $\Phi(\cdot)$ is the cumulative distribution function of a standard Gaussian distribution;
3. $g(t) = \text{arctanh}(2t/\delta - 1), \quad t \in (0,\ \delta), \quad g(t) \in (-\infty,\ \infty).$

In Cederkvist et al. (2019), this CIF specification is modeled under a pairwise composite likelihood approach (Lindsay 1988; Varin, Reid, and Firth 2011).

Our contribution: a full likelihood analysis

For two competing causes of failure, a subject i, in the cluster j, in time t, we have

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

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

$$\begin{aligned} p_{kijt} &= \frac{\partial}{\partial t} F_k(t \mid \boldsymbol{x}, \boldsymbol{u}, \eta_k) \\ &= \frac{\exp\{\boldsymbol{x}_{kij}\beta_k + u_{kj}\}}{1 + \sum_{m=1}^{K-1} \exp\{\boldsymbol{x}_{mij}\beta_m + u_{mj}\}} \\ &\times w_k \frac{\delta}{2\delta t - 2t^2} \phi \left(w_k \text{arctanh} \left(\frac{t - \delta/2}{\delta/2} \right) - \boldsymbol{x}_{kij}\gamma_k - \eta_{kj} \right), \quad k = 1,\ 2. \end{aligned} \tag{2}$$

Marginal likelihood function for two competing causes

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

with p_{kijt} from Equation 2 and where $\boldsymbol{\theta} = [\boldsymbol{\beta}\ \boldsymbol{\gamma}\ \boldsymbol{w}\ \boldsymbol{\sigma}^2\ \boldsymbol{\rho}]^{\top}$ is the parameters vector.

TMB: Template Model Builder

Kristensen et al. (2016).

An R (R Core Team 2021) package for the quickly implementation of complex random effect models through simple C++ templates.

Key features:

1. Automatic differentiation;
   The state-of-art in derivatives computation
2. Laplace approximation.
   An efficient fashion to approximate the latent effect integrals

Simulation study results

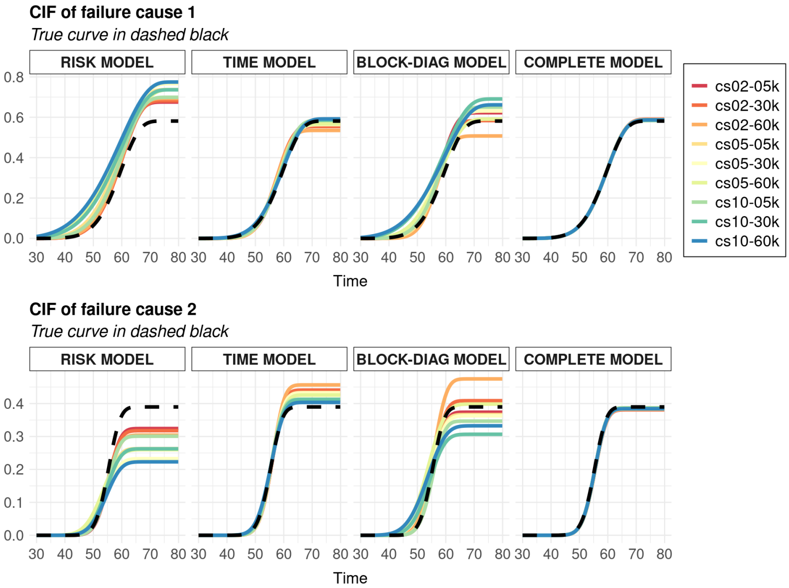
First of all, the **time**.

- In the most expensive scenarios (30K 4D Laplaces),
  the complete model takes 30 min.
  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.

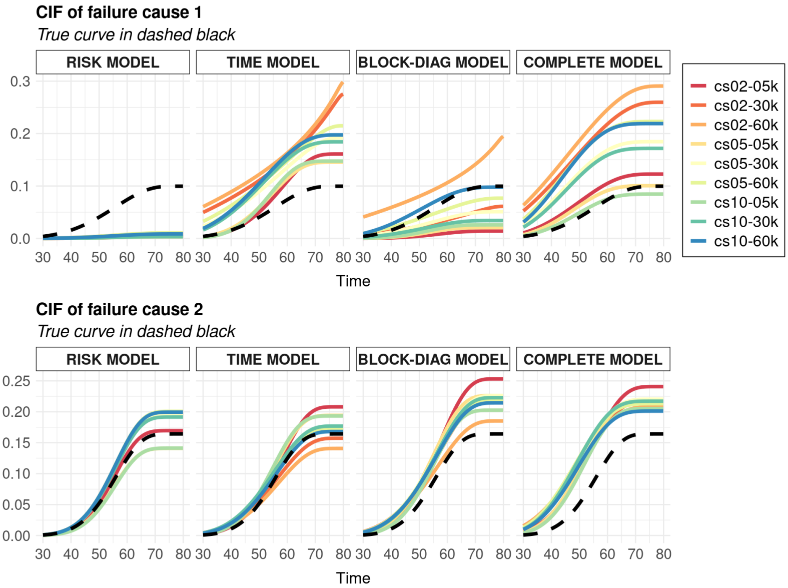
Parameters estimation.

- The *non-complete* models fail to learn the data.
  They appear to be *not structured enough* to capture the data characteristics.

Simulation study results: High CIF scenario



Simulation study results: Low CIF scenario



Thanks for watching and have a great day

For more read Laureano (2021) master thesis.

Special thanks to

PPGMNE
Programa de Pós-Graduação em
Métodos Numéricos em Engenharia

Joint work with

Wagner H. Bonat
http://leg.ufpr.br/~wagner

Paulo Justiniano Ribeiro Jr.
http://leg.ufpr.br/~paulojus

henriquelaureano.github.io

References

Cederkvist, L., K. K. Holst, K. K. Andersen, and T. H. Scheike. 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.

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

Laureano, H. A. 2021. "Modeling the Cumulative Incidence Function of Clustered Competing Risks Data: A Multinomial Glmm Approach." Master's thesis, Federal University of Paraná (UFPR).

Lindsay, B. G. 1988. "Composite Likelihood Methods." *Comtemporary Mathematics* 80 (1): 221–39.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Stan Development Team. 2020. "RStan: The R Interface to Stan." https://mc-stan.org/.

Varin, C., N. Reid, and D. Firth. 2011. "An Overview of Composite Likelihood Methods." *Statistica Sinica* 21 (1): 5–42.