A multinomial generalized linear mixed model for clustered competing risks data

Henrique Aparecido Laureano* Wagner Hugo Bonat*

August 16, 2021

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 the risk and failure time trajectory levels are correlated.

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

*Laboratório de Estatística e Geoinformação, Departamento de Estatística, Universidade Federal do Paraná, Curitiba, Brasil. E-mail: laureano@ufpr.br

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. The modeling object is the survival experience of the time-to-event data. 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.

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, described in Figure 1 and the focus of this work. More specifically, its clustered version i.e., with groups of elements sharing some non-observed latent dependence structure.

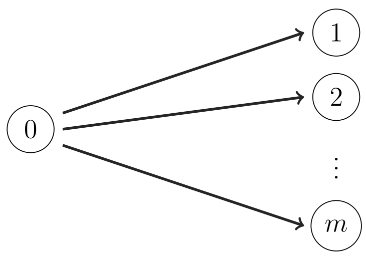

Figure 1: Illustration of competing risks process.

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.

In its simplest form, it is assumed that all subjects under study are independent. However, if the subjects are structured in a related fashion the independence assumption is unrealistic. When the subjects are organized in clusters (a family, e.g.), the nonindependence is accommodated in terms of a latent/random-effect, shared by all elements of that cluster. This idea opens space to what is called *family studies*. In family studies, the goal is to accommodate 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 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 instead with the of 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 to model the survival experiences in 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.

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 analysis. 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, easily extend to any number of competing causes. The within-cluster dependence is accommodated via a multivariate normal distribution and the cause-specific CIFs via the model's link function. The estimation and inference are done via an efficient implementation and state-of-art computational libraries provided through TMB. The latent effects are handled out by means of an efficient Laplace approximation and automatic differentiation.

The main goal of this study 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 the practical relevance of our framework. The estimation and inference is made through the efficient computational resources of the R (R Core Team; 2021) package `TMB` (Kristensen et al.; 2016).

The main contributions of this article are: (i) introducing the modeling of cause/cluster-specific CIFs of clustered competing risks data into an efficient implementation 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

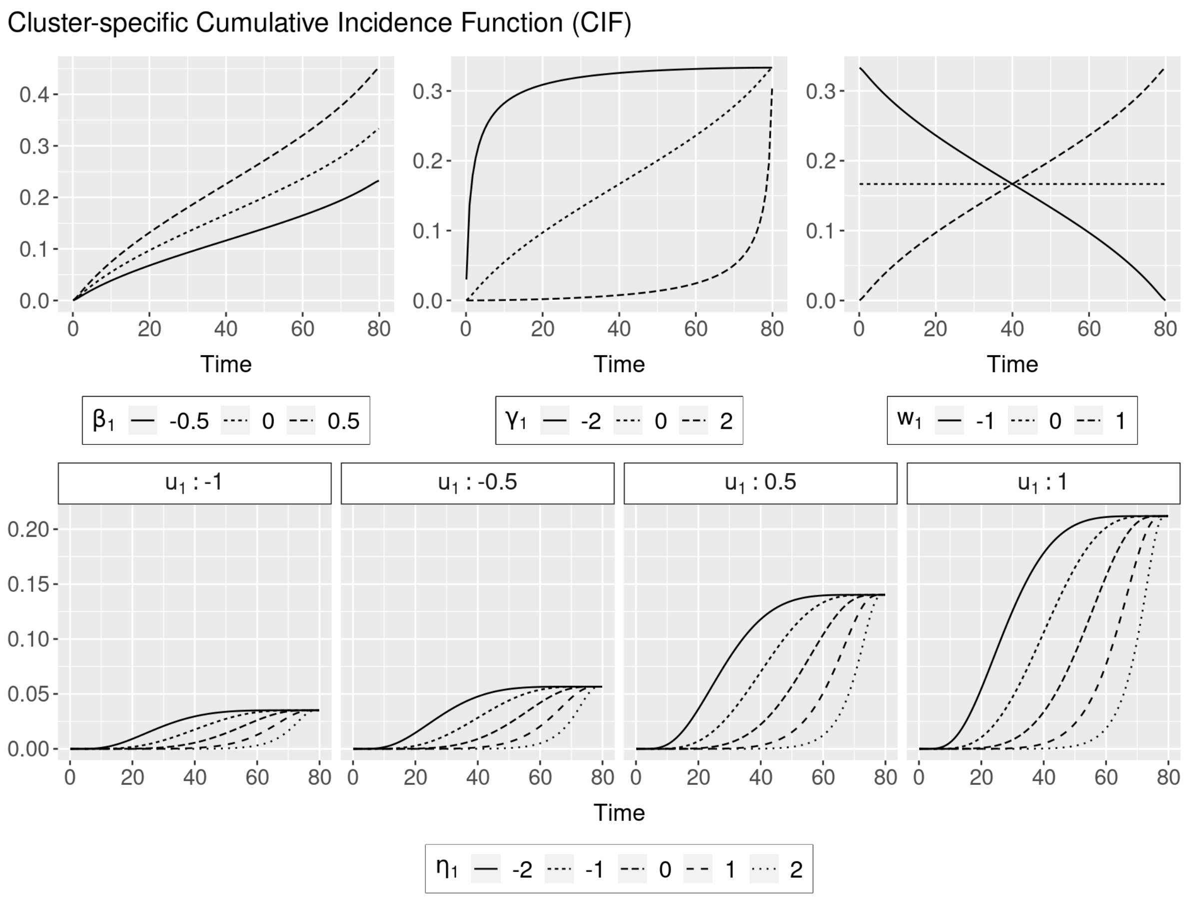

Figure 2: Curve behaviors for different parameter settings, showing then the corresponding parameter effects in a cluster-specific cumulative incidence function (CIF).

3 Estimation and inference

4 Simulation studies

5 Discussion

Supplementary material

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.

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.

Diggle, P., Heagerty, P., Liang, K. Y. and Zeger, S. (2002). *Analysis of Longitudinal Data*, second Edition edn, Oxford University Press, United Kingdom.

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.

Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. (2008). *Longitudinal Data Analysis*, CRC Press.

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.

Masarotto, G. and Varin, C. (2012). Gaussian copula marginal regression, *Electronic Journal of Statistics* **6**(1): 1517–1549.

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.

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

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.

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

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.

Verbeke, G. and Molenberghs, G. (2001). *Linear Mixed Models for Longitudinal Data*, Springer Series in Statistics, New York.