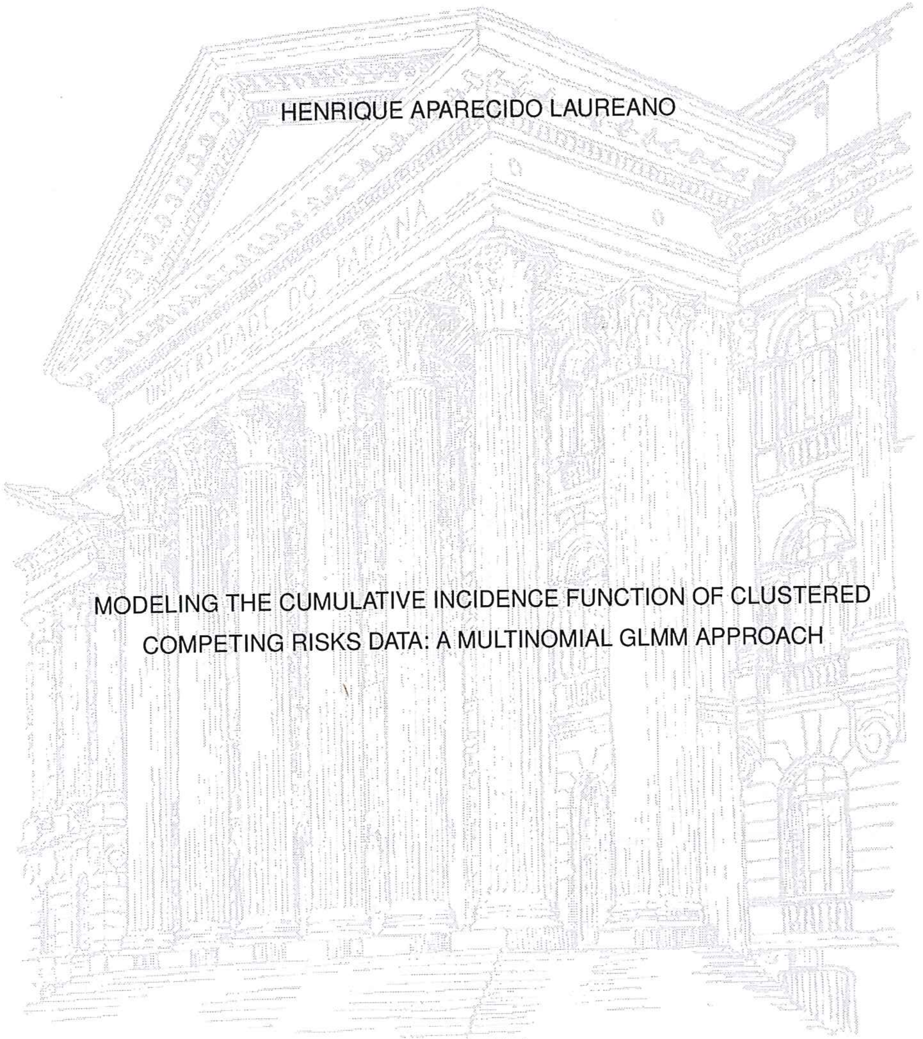
FEDERAL UNIVERSITY OF PARANÁ



HENRIQUE APARECIDO LAUREANO

MODELING THE CUMULATIVE INCIDENCE FUNCTION OF CLUSTERED COMPETING RISKS DATA: A MULTINOMIAL GLMM APPROACH

CURITIBA

2021

HENRIQUE APARECIDO LAUREANO

MODELING THE CUMULATIVE INCIDENCE FUNCTION OF CLUSTERED COMPETING RISKS DATA: A MULTINOMIAL GLMM APPROACH

Thesis presented to the Graduate Program of Numerical Methods in Engineering, Concentration Area in Mathematical Programming: Statistical Methods ~~Applied in Engineering~~, Federal University of Paraná, as part of the requirements to the obtention of the Master's Degree in Sciences.

Supervisor: Prof. PhD Wagner Hugo Bonat
Co-supervisor: Prof. PhD Paulo Justiniano Ribeiro Jr

CURITIBA

2021

HENRIQUE APARECIDO LAUREANO

MODELING THE CUMULATIVE INCIDENCE FUNCTION OF CLUSTERED COMPETING RISKS DATA: A MULTINOMIAL GLMM APPROACH

Thesis presented to the Graduate Program of Numerical Methods in Engineering, Concentration Area in Mathematical Programming: Statistical Methods ~~Applied in Engineering~~, Federal University of Paraná, as part of the requirements to the obtention of the Master's Degree in Sciences.

Master thesis approved. XXX XX, 2021.

Prof. PhD Wagner Hugo Bonat
Supervisor

Prof. PhD Paulo Justiniano Ribeiro Jr
Co-supervisor

Prof. PhD . . .
Internal Examinator - PPGMNE

Prof. PhD . . .
Internal Examinator - PPGMNE

Prof. PhD . . .
External Examiner -

CURITIBA
2021

To Celita and Olivio

ACKNOWLEDGEMENTS

As Moro once said, I am thankful for everything and everyone.

"It's not supposed to be easy."
(Gregg Popovich)

ABSTRACT

Clustered competing risks data is a special case of failure time data. Besides the cluster structure which implies a latent within-cluster dependence between its elements, this kind of data is characterized by 1) multiple causes/variables competing to be the one responsible for the occurrence of an event, a failure; and 2) censorship, when the event of interest happens or not for none of the competing causes, in the study period. To handle this type of data, we propose a generalized linear mixed model (GLMM) i.e., a latent-effects framework, instead of a usual survival model. In survival analysis, the modeling is usually done by means of the hazard rate, and the within-cluster dependence accommodation ends by generating a complicated likelihood function, sometimes intractable. We, on the other hand, model the clustered competing causes in the probability scale, in terms of the cumulative incidence function (CIF) of each competing cause. In our framework, we suppose a multinomial probability distribution for the competing causes and censorship, conditioned on the latent effects. The latent effects are accommodated via a multivariate Gaussian distribution and are modeled by the parameters of its covariance matrix. The probability distributions are connected via CIF, modeled here following Cederkvist et al. (2019) specification, based on its decomposition as the product of an instantaneous risk level function with a trajectory time level function. The latent effects are inserted in both ~~those~~ level functions. To make the model parameters estimation the most efficient as possible, we use the template model builder (TMB) (KRISTENSEN et al., 2016). With this R (R Core Team, 2021) package, we have 1) the log-likelihood function written in C++; 2) access to efficient linear algebra libraries; 3) efficient Laplace approximation implementation for the latent-effects; and 4) an automatic differentiation (AD) routine, the state-of-the-art in derivatives computation. To check the model performance estimability ~~of our model~~ a large simulation study is performed, based on different latent structure formulations, with the aim to verify which one is the most adequate to real scenarios. The model presents to be of difficult estimation, with our results converging to a latent structure where the risk and trajectory time levels are correlated. In scenarios with high CIF the model exhibits the best ~~better~~ results, but still with an excessive variance, showing that improvements are necessary.

Achei essa frase confusa.

não é a estimabilidade do modelo que está sendo avaliada.

Tenta evitar esse tipo de afirmação. O que vc entende por melhor ou pior não é algo preciso. Explique o pq vc achou que é melhor.

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

RESUMO

Dados de riscos competitivos agrupados são um caso especial de dados de tempo de falha. Além da estrutura de grupo que implica uma dependência latente intra-grupo entre seus elementos, esse tipo de dado é caracterizado por 1) múltiplas causas/variáveis competindo para ser a responsável pela ocorrência de um evento, uma falha; e 2) censura, quando o evento de interesse ocorre ou não por nenhuma das causas concorrentes, no período de estudo. Para lidar com este tipo de dado, propomos um modelo linear generalizado misto (GLMM), ou seja, um modelo de efeitos latentes/aleatórios, em vez de um modelo de sobrevivência usual. Em análise de sobrevivência, a modelagem é usualmente feita por meio da taxa de risco, e a acomodação da dependência intra-grupo acaba por gerar uma complicada função de verossimilhança, às vezes intratável. Nós, por outro lado, modelamos as causas competidoras agrupadas na escala da probabilidade, por meio da função de incidência acumulada (CIF, em inglês) de cada causa competidora. Em nossa modelagem, supomos uma distribuição de probabilidade multinomial para as causas competidoras e censura, condicionado aos efeitos latentes. Os efeitos latentes são acomodados por meio de uma distribuição Gaussiana multivariada e são modelados via os parâmetros de sua matriz de covariância. As distribuições de probabilidade são conectadas por meio da CIF, modeladas aqui seguindo a especificação em Cederkvist et al. (2019), com base em sua decomposição como o produto de uma função de nível de risco instantâneo com uma função de nível de tempo de trajetória. Os efeitos latentes são inseridos nestas funções. Para tornar a estimativa dos parâmetros do modelo o mais eficiente possível, usamos o template model builder (TMB) (KRISTENSEN et al., 2016). Com este pacote R (R Core Team, 2021), temos 1) a função de log-verossimilhança escrita em C++; 2) acesso a eficientes bibliotecas de álgebra linear; 3) implementação eficiente da aproximação de Laplace para os efeitos latentes; e 4) uma rotina computacional de diferenciação automática, o estado da arte em computação de derivadas. Para verificar a performance estimabilidade do ~~nosso~~ modelo foi é realizado um amplo estudo de simulação, baseado em diferentes formulações de estruturas latentes, com o objetivo de verificar qual delas é a mais adequada a um cenário real. O modelo se apresenta de difícil estimatição, com nossos resultados convergindo para uma estrutura latente onde os níveis de risco e de trajetória estão correlacionados. Em cenários de CIF alta o modelo apresenta os melhores resultados, mas ainda com uma excessiva variablidade, mostrando que melhorias são necessárias.

Achei confusa

Palavras-chave: Riscos competitivos agrupados. Dependência intra-cluster. Modelo linear generalizado misto multinomial (MLGM). TMB: Template Model Builder. Aproximação de Laplace. Diferenciação automática.

LIST OF FIGURES

LIST OF ALGORITHMS

CONTENTS

1 INTRODUCTION

Consider a cluster of random variables representing the time until the occurrence of some event. These random variables are assumed to be correlated, i.e. for some biological or environmental reason it is not adequate to assume independence between them. Also, we may be interested in the occurrence of not only one specific event, having in practice a competition of events to see which one happens first, if it happens. Such events may also be of low probability albeit severe consequences, this is the moment when the cluster correlation makes its difference: the occurrence of an event in a cluster member should affect the probability of the same happening in the others.

A realistic context that fits perfectly with the framework described above is the study of disease incidence in family members, where each member is indexed by a random variable and each cluster consists of a familiar structure. The inspiration to ~~the~~ investigate ~~study of~~ these kinds of problems came from the work developed in Cederkvist et al. (2019), where they studied breast cancer incidence in mothers and daughters ~~but using~~ a complicated modeling framework. Based on that, the aim of this thesis is to propose a simpler framework taking advantage of several *state-of-art* computational libraries and see how far we can go in several scenarios. Until now we have just contextualized, we still need to introduce the methodology. To do this, some definitions and theoretical contexts are welcome.

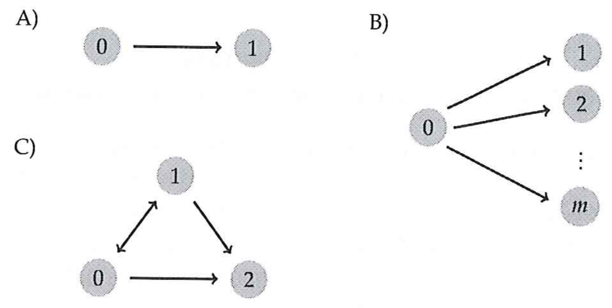
Vc usa o mesmo modelo só muda o esquema de estimação. Aqui parece que vc usa outro modelo

When the object under study is a random variable representing the time until some event occurs, we are in the field of *failure time data* (KALBFLEISCH; PRENTICE, 2002). The occurrence of an event is generally denoted *failure*, and major areas of application are biomedical studies and industrial life testing. In this thesis, we maintain our focus on the former. no último? As common in science, same methodologies can receive different names depending on the area. In industrial life testing is performed what is called a *reliability analysis*; in biomedical studies is performed what is called *survival analysis*. Generally, the term *survival* is applied when we are interested in the occurrence of only one event, a *failure time process*. When we are interested in the occurrence of more than one event we enter in the yard of *competing risks* and *multistate* models. A visual aid is presented on Figure 1 and a comprehensive reference is Kalbfleisch & Prentice (2002).

Failure time and competing risks processes may be seen as particular cases of a multistate model. Besides the number of events (states) of interest, the main difference between a multistate model and its particular cases is that only in the multistate scenario we may have transient states, using a *stochastic process* language. In the particular cases, all states besides the initial state 0, are absorbents - once you reached it you do not leave.

The simplest multistate model that exemplify this behavior is the illness-death model, Figure 1 C), where a patient (initially in state 0) can get sick (state 1) or die (state 2); if sick it can recover (returns to state 0) or die. We work in this thesis only with competing risks processes, and for each patient we need the time (age) until the occurrence, or not, of the event.

FIGURE 1 – ILLUSTRATION OF MULTISTATE MODELS FOR A A) FAILURE TIME PROCESS; B) COMPETING RISKS PROCESS; AND C) ILNESS-DEATH MODEL, THE SIMPLEST MULTISTATE MODEL



SOURCE: The author (2021).

When for some known or unknown reason we are not able to see the occurrence of an event, we have what is denoted *censorship*. Still in the illness-death model, during the period of follow up the patient may not get sick or die, staying at state 0. This is denoted *right-censorship*; if a patient is in state 1 at the end of the study, we are *censored* to see him reaching the state 2 or returning to state 0. This is the inherent idea to censorship and must be present in the modeling framework, thus arriving in the so-called *survival models* (KALBFLEISCH; PRENTICE, 2002).

A survival model deals with the survival experience. Usually, the survival experience is modeled in the *hazard* (failure rate) scale and it can be expressed for a subject i as

$$\lambda(t \mid x_i) = \lambda_0(t) \times c(x_i\beta) \quad \text{at time } t, \tag{1.1}$$

i.e. as the product of an arbitrary baseline hazard function $\lambda_0(\cdot)$, with a specific function form $c(\cdot)$, that will depend on the probability distribution to be chosen for the failure time and on predictors/covariates/explanatory/independent variables $x_i = [x_1 \ \dots \ x_p]$, where $\beta^\top = [\beta_1 \ \dots \ \beta_p]$ is the parameters vector.

This structure is specified for a failure time process, as in Figure 1 A). Nevertheless, the idea is easy to extend. We basically have the Equation 1.1's model to each cause-specific (in a competing risks process) or transition (in a multistate process). A complete and extensive detailing can be, again, found in Kalbfleisch & Prentice (2002).

In this work we approach the case of clustered competing risks. Besides the cause-specific structure, we have to deal with the fact that the events are happening in related individuals. This configures what is denoted *family studies*, i.e. we have a cluster/group/family dependence that needs to be considered, accommodated, and modeled. This, possible, dependence is something that we do not actually measure but know (or just suppose) that exists. In the statistical modeling language this characteristic receives the name of *random* or *latent effect*. A survival model with a latent effect, association, or unobserved heterogeneity, is denoted *frailty model* (CLAYTON, 1978; VALPEL; MANTON; STALLARD, 1979). In its simplest form, a frailty is an unobserved random proportionality factor that modifies the hazard function of an individual, or of related individuals. Frailty models are extensions of Equation 1.1's model.

In the competing risks setting, the hazard scale (focusing on the cause-specific hazard) is not the only possible scale to work on. A more attractive possibility is to work on the probability scale (ANDERSEN et al., 2012), focusing on the cause-specific cumulative incidence function (CIF). Besides the within-family dependence, in family studies there is often a strong interest in describing age at disease onset, which is directly described by the cause-specific CIF. Therefore, making the probability scale a more attractive and logical choice. Since the CIF plays a central role in this master thesis, it will be formally defined later in a place with greater emphasis. With the definitions and the theoretical context being made, let us be more specific.

To work with competing risks data on the probability scale plus a latent structure allowing for within-cluster dependence of both risk and timing, Cederkvist et al. (2019) proposed a pairwise composite likelihood approach based on the factorization of the cause-specific CIF as the product of a cluster-specific risk level function with a cluster-specific failure time trajectory function. A composite approach (LINDSAY, 1988; COX; REID, 2004; VARIN; REID; FIRTH, 2011) is a valid alternative to a full likelihood analysis in high-dimensional situations when a full approach is too computational costly or even inviable. In failure time data problems, the composite likelihood function is built from the product of marginal densities. The marginal specification implies a pairwise approach since we need to add model layers to be able to handle with the dependence structure. A clear advantage of this approach is that we do not need to care about a joint distribution specification, which generally translates also into a computational advantage. A disadvantage is the model specification, which becomes much more complicated, besides the number of small details to workaround from the fact of being working with not an exact likelihood function.

O modelo é o mesmo, o que vc vai mudar é o esquema de estimação. Isso não está claro aqui!

Não sei se entendi o que vc quer dizer, mas vejo que vc está falando de especificação do modelo e confundindo com o método de estimação

We do not have any guarantees that a full likelihood inference procedure is not viable here, so we try to reach the same goal of Cederkvist et al. (2019) albeit with a simpler framework taking advantage of *state-of-art* software, something still not so common in the statistical modeling community. This simpler framework is a generalized linear mixed model (GLMM). Instead of concentrating on failure time data and consequently having a survival/frailty model based on the hazard scale, or using a composite approach, we just build the joint/full likelihood function (a multinomial model with its link function based on the cluster-specific CIF, accouting for an appropriate latent effects structure), marginalize (integrate out the latent effects) and optimize it. A Fisherian approach per se.

Não entendi o que vc quer com isso?

To a better contextualization of our GLMM approach (MCCULLOCH; SEARLE, 2001), consider a random suject i. In a standard linear model we assume that the response variable Y_i, conditioned on the covariates x_i, follows a normal/Gaussian distribution and what we do is to model its mean, $\mu_i \equiv \mathbb{E}(Y_i \mid x_i)$, via a linear combination. As much well explained in Nelder & Wedderburn (1972), with the aid of a *link function* $g(\cdot)$, this idea is generalized to distributions of the *exponential family*. Many of its members are useful for practical modelling, such as the Poisson (for counting data), binomial (dichotomic data), gamma (continuous but positive) and Gaussian (continuous data) distributions. This extended framework received the name of generalized linear model (GLM) and a comprehensive reference is McCullagh & Nelder (1989).

What makes a GLM into a GLMM (MCCULLOCH; SEARLE, 2001) is the addition of a latent effect u (then, *mixed*) into the mean structure. The mean structure of a standard GLMM is defined as

$$g(\mu_i) = x_i\beta + z_i u, \quad u \sim \text{Multivariate Normal}(\mathbf{0}, \mathbf{\Sigma})$$

where the latent effect is assumed to follow a multivariate Gaussian distribution of zero mean and a parametrized variance-covariance matrix $\mathbf{\Sigma}$. Its correct linkage to the mean structure is made through the i^{th} vector row of a design-matrix $\mathbf{Z}$. The covariates are into x_i, the i^{th} vector row of a model-matrix $\mathbf{X}$, with $\boldsymbol{\beta}$ being a vector of unknown parameters.

1.1 GOALS

1.1.1 General goals

performance of maximum likelihood estimators

Propose and study the estimability of a multinomial generalized linear mixed model (multiGLMM) to the cluster and cause-specific cumulative incidence function (CIF) of clustered competing risks data.

1.1.2 Specific goals

1. Simulate from the model, i.e. generate synthetic data to study statistical properties.

2. Write the model in the Template Model Builder (TMB) software, developed by Kristensen et al. (2016) and possibly the most efficient likelihood-based way of doing such task.
3. Take advantage of TMB's functionalities with special attention to the computation of gradients and Hessians via a *state-of-art* automatic differentiation (AD) implementation; and a joint likelihood marginalization via an efficient Laplace approximation routine.
4. Study the model identifiability through the proposition of different complexity level models in terms of parametric space and latent effect structures.
5. Make exact likelihood-based inference to the cluster and cause-specific CIF of clustered competing risks data.

1.2 JUSTIFICATION

In the biomedical statistical modeling literature, the study of disease occurrence in related individuals receives the name of family studies. Key points of interest are the within-family dependence and determining the role of different risk factors. The within-family dependence may reflect both disease heritability and the impact of shared environmental effects. The role of different risk factors arrives in the class of multivariate models, which options are limited in the statistical literature. Thus, the number of statistical models for competing risks data that accommodate the within-cluster/family dependence is even more limited. Some modeling options are briefly commented in Cederkvist et al. (2019), with his pairwise composite approach being proposed as a new and better option to model the cause-specific cumulative incidence function (CIF), describing age at disease onset, of clustered competing risks data on the probability scale. We propose to model the cause-specific CIF and accommodate the within-family dependence in the same fashion (via a latent structure that allows the absolute risk and the failure time distribution to vary between families) but with an easier framework, based on a multinomial generalized linear mixed model approach.

1.3 LIMITATION

não!

This work restraint to the proposition and model identifiability study of a multinomial model for the cause-specific cumulative incidence function (CIF) of competing risks data, with a latent effect structure to accommodate within-family dependence with regard to both risk and timing. Given its considerable model complexity, hypothesis tests; residual analysis; and good-of-fit measures are not contemplated.

Inclua que vc se restringe a family studies onde em geral ~~os~~ o número de clusters é grande porém elementos dentro de cada cluster é pequeno.

1.4 THESIS ORGANIZATION

This master thesis contains 6 chapters including this introduction. Chapter 2 presents a systematic review of the main aspects involved in the formulation, optimzation, and implementation of a generalized linear mixed model (GLMM). Given the modeling framework overview, Chapter 3 presents our multinomial GLMM (multiGLMM) to model the cause-specific cumulative incidence function (CIF) of clustered competing risks data. In Chapter 4 we describe the simulation procedure to generate synthetic data and present some model particularities. In Chapter 5 the obtained results are presented, and in Chapter 6 we discuss the contributions of this thesis and present some suggestions for future work.

Vc está estudando coisas relacionadas ao método de estimação não necessariamente sobre o modelo. Mesmo que vc chegasse a conclusão que vc não conseguiu estimar seria o que usando o método de máx verossimilhança vc não consegui estimar. Isso não significa que o modelo não é estimável.

Vc precisa alterar todo a sua motivação para explicar que vc está avaliando propriedades do esquema de estimação e não do modelo