Joint models: Scientific initiation report

Rodrigo Kalil, Luiz Max Carvalho
School of Applied Mathematics, Getulio Vargas Foundation, Brazil.

March 29, 2025

Abstract

The analysis of clinical data often involves both measurements taken during visits over time, known as longitudinal measurements, and data indicating whether an event has occurred, referred to as time-to-event data. Analyzing these variables separately has limitations, which joint models aim to overcome. In a joint model, we consider both types of variables and estimate parameters associated with them. This allows us to analyze their correlation and how they influence each other.

This project aims to study this class of models, which has been developed over the past 25 years, and to expand existing implementations available in R packages. Additionally, if possible, apply these packages to epidemiological data.

Key-words: joint models; longitudinal analysis; survival analysis; stan; jmbayes2.

Activities

Process nº 201.340/2024
Start (31/01/2024) End (31/01/2025)

- **February, 2024** - Reading of Henderson's article.

  Joint modelling of longitudinal measurements and event time data was the first paper to introduce the concept of joint models, already immersed in medical contexts. Studying it deeply was important to understand the theory behind it's formulation and the essential methods to the implementation. The paper also thought the motivation to this models in clinical applications.

- **March, 2024** - Reading and technical review of Damone's article.

  This paper brings the state of art in term of joint modeling. Following the novelties and approaches to deal with yet limitations on this field was an important step during the project. Because it was a preparation to future applications of the models. Besides that, Damone's a collaborator that could help in future work.

- **From April to June, 2024** - Review of the code intending to replicate Henderson's implementation

  There was an effort of other student, Ezequiel Braga, who previously simulated standard models with R and Stan. The review of this code was important to get used to statistic modeling in such languages. Moreover, the simulation of Henderson's scenario would be a checkpoint to assure consistency in the code.

- **From April to June, 2024** - Review of avaiable software

  We found difficulty on replicating the implementation of joint models, This motivated the reviewing of software with support to joint models, allowing a better understanding of the state of art in terms of computational tools for this field.

- **From August to November, 2024** - Numerical stability study of interest integral

  The comprehension of the code and model showed problems with the implemantations. In special, there's an integral important for the implementation that in theory would assume problematic values in certain conditions. Therefore, a study was made and it was found that indeed some regions of parameter space presented extreme values.

Background

This section provides the theoretical foundation necessary to understand joint models, given on Henderson et al. (2000)'s article, covering the key mathematical and statistical components involved.

Why Joint Models?

Joint models are used to analyze longitudinal and survival data simultaneously. They allow us to account for time-dependent co-variates and association structures that traditional separate models fail to capture. The main motivation behind joint modeling is to improve predictive accuracy and better understand the relationship between longitudinal biomarkers and survival outcomes.

What are Joint Models?

Joint models combine longitudinal and survival sub-models. In Henderson's they are called measurements and intensity sub-models and are composed via a zero-mean bi-variate Gaussian process, $W_i(t) = \{W_{1i}(t), W_{2i}(t)\}$.

For the ith subject, the measurements Y at times t are given as follows.

$$Y_{ij} = \mu_i(t_{ij}) + W_{1i}(t_{ij}) + Z_{ij} \tag{1}$$

where:

- $\mu_i(t_{ij})$ is the mean response
- $Z_{ij} \sim N(0, \sigma_Z^2)$ is a sequence of independent errors

It's assumed the mean response may be modeled as follows.

$$\mu_i(t) = x_{1i}(t)^T \beta_1 \tag{2}$$

where:

- $x_{1i}(t)$ represents possibly time-varying explanatory variables
- β_1 is it's respective coefficients

On the other hand, the event intensity is modeled as follows.

$$\lambda_i(t) = H_i(t)\alpha_0(t)exp\{x_{2i}(t)^T \beta_2 + W_{2i}(t)\} \tag{3}$$

where:

- $\alpha_0(t)$ is unspecified
- $x_{2i}(t)$ and β_2 are the explanatory variables with their coefficients, which may have itens in common with the longitudinal ones

The combination of the measurements and intensity formulations above includes a great range of specific models previous used to separate analysis. So, it's possible to reuse formulations already proposed and tested for longitudinal and event values.

It is indeed possible to choose separately the structure for each sub-model and than combine them as stated. In particular, here is a combination proposed for the longitudinal part.

$$W_{1i}(t) = d_{1i}(t)^T U_{1i} + V_{1i}(t) \tag{4}$$

where:

- $d_{1i}(t)$ is the vector of explanatory variables
- U_{1i} is a vector of random effects distributed by a multi-variate normal
- $V_{1i}(t)$ is a stationary Gaussian process

How are Joint Models Applied?

Joint models are commonly used in medical research, particularly in analyzing biomarkers and disease progression. They are implemented using Bayesian and frequentist approaches, leveraging software such as JMbayes2 in R or Stan. The estimation process typically involves a combination of maximum likelihood estimation (MLE) or Markov Chain Monte Carlo (MCMC) methods to account for the complexities of the joint likelihood function.

Reviewing the article "A Bayesian Approach for Joint Models of Recurrent Events, Terminal Events and Longitudinal Data"

According to Damone's article, joint models are usual both on models with recurrent and terminal events and on models with longitudinal data and terminal events. But there is a lack of models able to deal with the three kinds. Besides that, the ones proposed on this use strong correlation and no bayesian approach.

The article proposes a model using random effects to connect a survival model - including recurrent and terminal - with a longitudinal outcome model. Proportional hazard models are used to model dependence between recurrent and terminal events, while a set of random but correlated effects is used in a linear mixed model to model dependence with longitudinal outcome measures. The random effects are connected by a multivariate normal.

The model is tested with the ARIC study data(atherosclerosis risk in communities). In this study, about 16000 individuals between 45 and 65 years old and placed among four communities in the USA were followed. From 1987 to 2019, 7 visits occurred, with a great pause before the fifth visit, on which data such as blood pressure and cholesterol was collected. Recurrent hospitalization of patients due to heart illness and death was also registered. The study concluded that systolic blood pressure is a risk factor for coronary arterial disease, and this data was already analyzed with other joint models previously.

The proposed model connects the submodels in such a way there's less direct association between the events and longitudinal data at the assumed parameters. In general, a covariance matrix would support one of the submodels, relating it with the others. However, the functions

of this model are set independent and random effects are correlated lately with a multivariate normal distribution.

Due to the nature of the longitudinal data, it is proposed a linear model to longitudinal data and an alternative piecewise model that allows a differentiation of data before and after the great period without measurement. The mWAIC metric is used to compare the two approaches and it was concluded that the alternative piecewise model performs better in this case.

Based on Henderson's work, where a covariance matrix links the submodels, the link alternative brought by Damone may be useful when correlationing the longitudinal submodels with others that are not of time-to-event type, yet such a path was not proposed.

The application of piecewise functions has no direct relation with the motivation of the scientific initiation, but is an interesting stratagem. The mWAIC and other metrics used when comparing mixed linear models, mainly used with longitudinal data, is explained at the referenced article "Bayesian model selection in linear mixed models for longitudinal data."

Existing software

- **JM** Rizopoulos (2010)

  The separate analysis of longitudinal and time-to-event data was possible using previous packages, but only with JM there's a computational tool to model jointly such variables. The package offers flexibility when defining the survival model, making it more useful. The remark, which stimulates the development of the next package, is that it estimates using maximum likelihood, while previous literature have used both maximum likelihood and a Bayesian approach with MCMC.

  Pros: Is the first package that allows the joint analysis of longitudinal and time-to-event variables.

  Cons: It supports a limited range of models, facing badly higher complexity scenarios. It is overcome by JMBayes.

- **JMBayes** Rizopoulos (2016)

  The package fits joint models with a Bayesian approach. In comparison to the JM package, from the same author, which uses maximum likelihood, JMBayes can fit a greater diversity of models. It also offers more tools to manipulate the predictions' results and quantify their's quality.

  The theoretical framework for the package is similar to the used by Henderson, assigning a longitudinal and a time-to-event functions. The estimation proceeds with a MCMC. B-splines are used to the survival process hazard, which allows flexibility.

  To solve the integral on the survival process, the package employees standard Gauss-Kronrod and Gauss-Legendre quadrature rules.

  Pros: Extends methods to a greater diversity of model specification.

  Cons: It still accepts only strict longitudinal outcomes. Is overcome by JMBayes2.

- **rstanarm**, the `jm` module Brilleman et al. (2018)

  This package allows to compute the estimation of the joint models in Stan. A great advantage in doing so is that Stan offers many post-estimation functionalities, as model diagnosis and posterior predictions.

Pros: Stan provides lots of estimation functionalities.

Cons: It doesn't have specific support for joint modeling.

- **INLAjoint** Rustand et al. (2023)

  The package rely on estimation based on the integrated nested Laplace approximation algorithm. Rustand uses this to estimate joint models with several variables and showed that it is less computationally expensive than approaches using maximum likelihood, which would demand high dimension integrals for many longitudinal and survival submodels.

  Pros: Brings an alternative for previous maximum likelihood approaches, which wouldn't succeed when facing high complexity problems.

  Cons: Is still limited to a maximum likelihood approach, without following the bayesian options of JMBayes.

- **JMBayes2** Rizopoulos et al. (2024)

  The latest package developed by Rizopoulos into joint modeling breaks the strictness of longitudinal outcomes, allowing different types and assuming different distributions.

  It's still being updated, having it's latest version released on 2025-02-27, with changes on modeling functions and new datasets options.

  Pros: Is more flexible than JMBayes, allowing better model specification.

  Cons: Is still being developed.

Code availabilty

The code for this study is available at Github: rckalil. It was forked from the previous work of Ezequiel Braga at Github: EzequielEBS. There are new archives related to the experiments run during this project, including an integral stability study and simple implementations with JMBayes2.

Integration of the hazard function: numerical stability study

Motivation

The code which aims to recreate Henderson's experiments presented inconsistent results. It may be caused by calculation errors at an integral. Reviewing the article, the integral form is identified.

$$I = \int_0^t u^{\gamma-1} \exp(\alpha\beta u) du \tag{5}$$

When $\alpha \cdot \beta < 0$, the result is equal to

$$\frac{\Gamma(\gamma)}{(|\alpha\beta|)^{\gamma}} \cdot F_G(t; \gamma; |\alpha\beta|), \tag{6}$$

where $F_G(\cdot; a; b)$ is the CDF of a Gamma with shape $a > 0$ and rate $b > 0$.

On the other hand, there's no closed formula when $\alpha \cdot \beta > 0$. Therefore, we believe there are miscalculation in Stan, which has been used.

Experiment

To check if there's instability at the calculation of such a function in stan, a R script was written, as for each combination of the parameters detailed at the table below, calculates the integral with standard R and extracts an error metric; calls a stan script that calculates the same integral; save this values and the difference between the integrals at a '.txt' archive.

	α	β	γ
start	-10	-10	0.1
end	10	10	10
quantity	10	10	5

Each parameter was tested in a certain quantity of values varying within the chosen limits. The integral was calculated within the limits of [0,1].

Results

The results are showed in a less detailed vision, but it comprises all the context. The views may be explored in more details at the repo.

Each heatmap contains alpha and beta varying at the vertical and horizontal axes, respectively, with fixed gamma. The red indicates values at the upper bound of the scale. The values presented correspond to the scientific importance for the metrics on each graphic.

Some instability appears mainly when alpha or beta are near to zero. This cases are already indicated by the error metric of R.

The main region of concern, where $\alpha \cdot \beta > 0$, is not a cause of miscalculation, as the standardized figure shows. The divergence of R and Stan results is due to enlargement of scale.

Discussion

The integral appears not to be susceptible to calculation errors. The cause of parameters estimation errors must rely on other issue.

Conclusion

Throughout this project, both classic and recent literature on joint models were studied. Several statistical software packages were also explored, evaluating their advantages. The use of the rstanarm and JMBayes2 packages was investigated. Given challenges in reproducing Henderson's results, an experiment was conducted to assess the stability of an integral of interest.

Although it was not possible to extend the study to the analysis of epidemiological data, all activities were valuable in introducing the student to the research environment. The literature review, experimentation, and interactions with the professor and other student researchers in the field contributed significantly to the student's academic development.

References

Brilleman, S., Crowther, M., Moreno-Betancur, M., Buros Novik, J., and Wolfe, R. (2018). Joint longitudinal and time-to-event models via Stan. StanCon 2018. 10-12 Jan 2018. Pacific Grove, CA, USA.

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480.

Rizopoulos, D. (2010). JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, 35(9):1–33.

Rizopoulos, D. (2016). The R package JMbayes for fitting joint models for longitudinal and time-to-event data using mcmc. *Journal of Statistical Software*, 72(7):1–45.

Rizopoulos, D., Papageorgiou, G., and Miranda Afonso, P. (2024). *JMbayes2: Extended Joint Models for Longitudinal and Time-to-Event Data*. R package version 0.5-0.

Rustand, D., van Niekerk, J., Krainski, E. T., Rue, H., and Proust-Lima, C. (2023). Fast and flexible inference for joint models of multivariate longitudinal and survival data using integrated nested laplace approximations. *Biostatistics*.

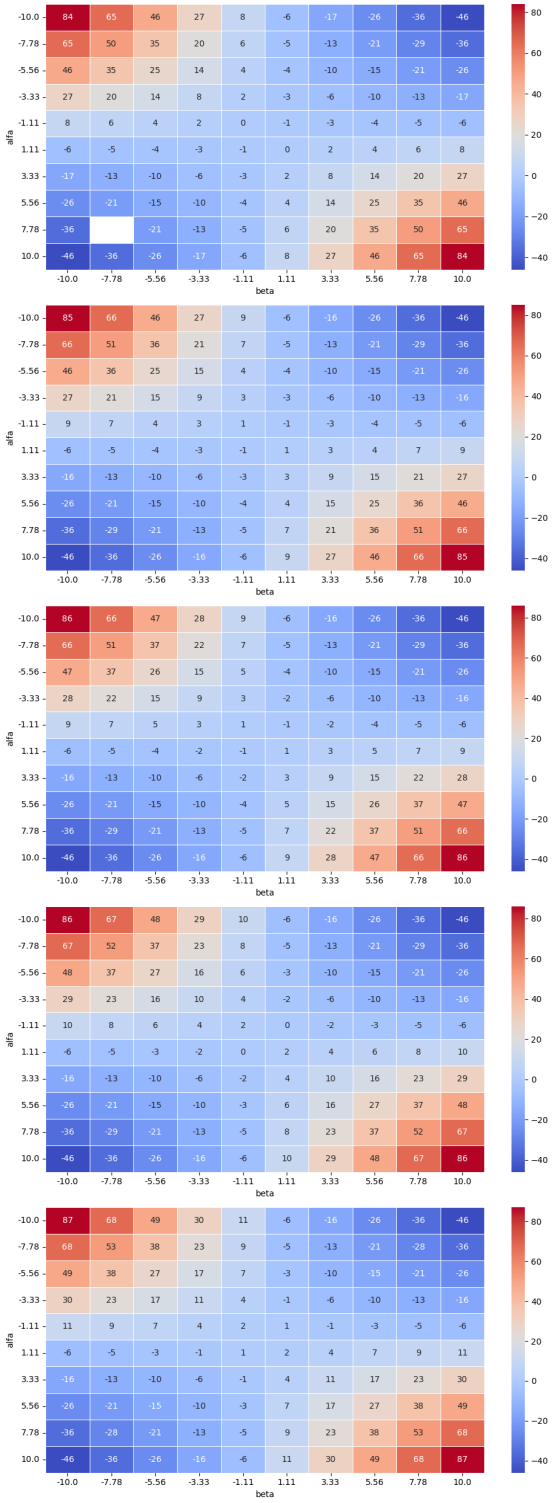

Figure S1: Area in R. It looks like a saddle.

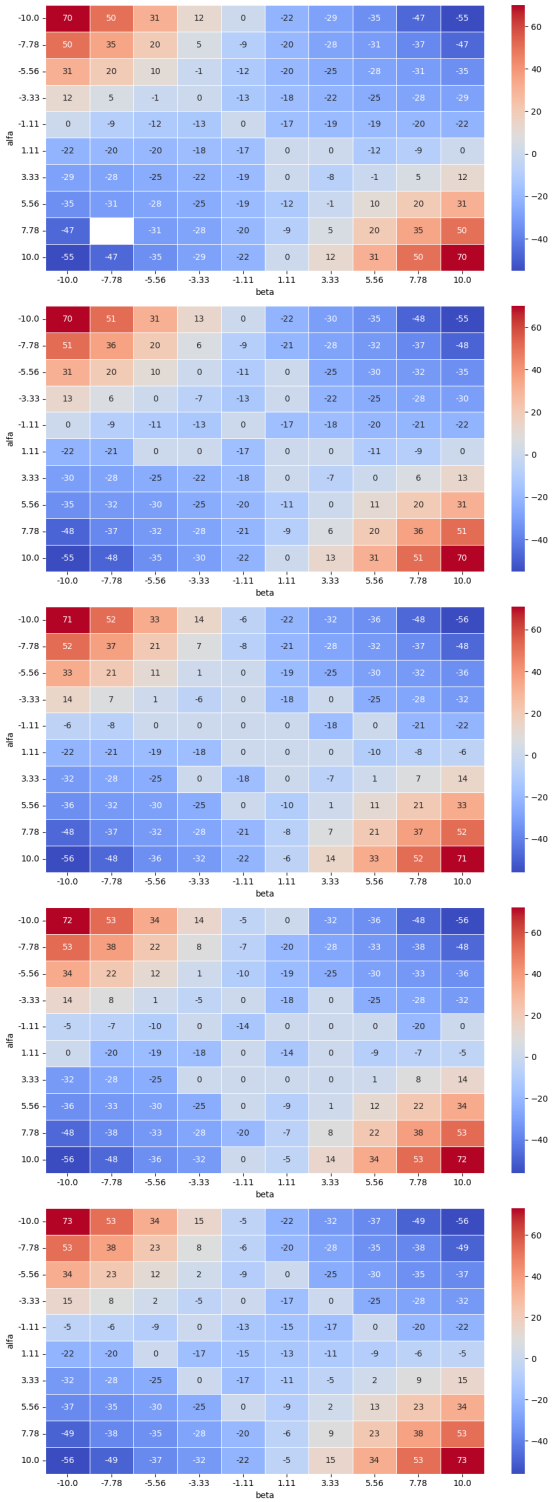

Figure S2: Difference between R and stan results. The values no long change softly, remarking enlargement of scale.

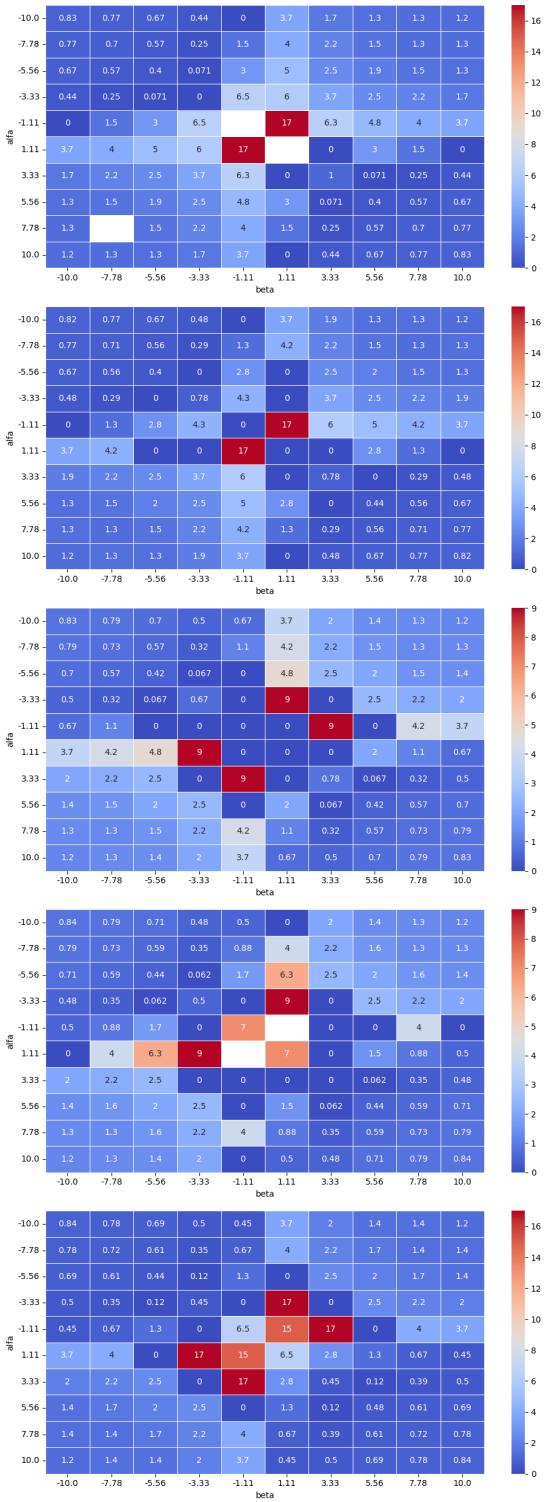

Figure S3: Difference standardized by R's area, in module. Red remarks the central region, where the integral gets close to zero.

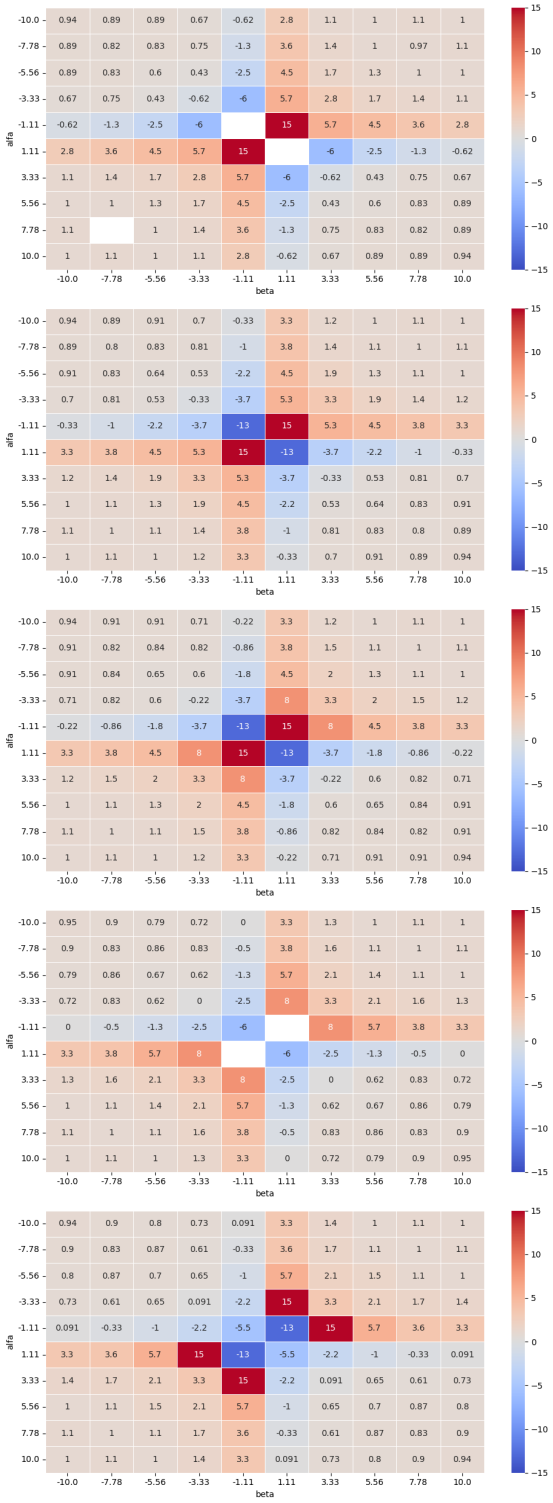

Figure S4: Quadrature error provided by R, related to the integral calculation.