Juliana Freitas de Mello e Silva

Joint Modeling Longitudinal and Survival Data via Bernstein Polynomials

(***Modelagem Conjunta de Dados Longitudinais e de Sobrevivência via Polinômios de Bernstein***)

Belo Horizonte, Brasil

Março de 2020

Juliana Freitas de Mello e Silva

Joint Modeling Longitudinal and Survival Data via Bernstein Polynomials

(*Modelagem Conjunta de Dados Longitudinais e de Sobrevivência via Polinômios de Bernstein*)

Tese apresentada ao Curso de de Pós-Graduação em Estatística da UFMG, como requisito para a obtenção do grau de DOUTORA em Estatística.

Universidade Federal de Minas Gerais – UFMG

Instituto de Ciências Exatas

Programa de Pós-Graduação em Estatística

Supervisor: Vinícius Diniz Mayrink

Co-supervisor: Fábio Nogueira Demarqui and Sujit Kumar Ghosh

Belo Horizonte, Brasil

Março de 2020

x

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa CRB 6ª Região nº 1510

Silva, Juliana Freitas de Mello e.

S586m Modelagem conjunta de dados longitudinais e de sobrevivência via polinômios de Bernstein / Juliana Freitas de Mello e Silva — Belo Horizonte, 2020.
120 f. il.; 29 cm.

Tese(doutorado) - Universidade Federal de Minas Gerais – Departamento de Estatísitca.

Orientador: Vinícius Diniz Mayrink.
Coorientador: Fábio Nogueira Demarqui

1. Estatística - Teses. 2. Análise de sobrevivência (Biometria). 3. Métodos de simulação. 4. Markov, Processos de 5. Método de Monte Carlo. 6. Estudos longitudinais. I. Orientador. II. Coorientador.III.Título.

CDU 519.2 (043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

ATA DA DEFESA DE TESE DA ALUNA

JULIANA FREITAS DE MELLO E SILVA

Realizou-se, no dia 27 de março de 2020, às 14:00 horas, por meio de sistema de vídeoconferência da Universidade Federal de Minas Gerais, a 62ª defesa de tese, intitulada *Modelagem Conjunta de Dados Longitudinais e de Sobrevivência via Polinômios de Bernstein*, apresentada por JULIANA FREITAS DE MELLO E SILVA, número de registro 2016671216, graduada no curso de ESTATÍSTICA, como requisito parcial para a obtenção do grau de Doutora em ESTATÍSTICA, à seguinte Comissão Examinadora: Prof. Vinícius Diniz Mayrink - Orientador (Depto. de Estatística / UFMG), Prof. Fábio Nogueira Demarqui - Coorientador (Depto. de Estatística / UFMG), Prof. Enrico Antônio Colosimo (Depto. de Estatística / UFMG), Prof. Dani Gamerman (Depto. de Estatística / UFMG), Profa. Leila Denise Alves Ferreira Amorim (Depto. de Estatística / UFBA) e Prof. Antônio Eduardo Gomes (Depto. de Estatística / UnB).

A Comissão considerou a tese:

(X) Aprovada

( ) Reprovada

Finalizados os trabalhos, lavrei a presente ata que, lida e aprovada, vai assinada por mim e pelos membros da Comissão.

Belo Horizonte, 27 de março de 2020.

Prof. Vinícius Diniz Mayrink - Orientador (Doutor)

Prof. Fábio Nogueira Demarqui - Coorientador (Doutor)

Prof. Enrico Antônio Colosimo (Doutor)

Prof. Dani Gamerman (Doutor)

Profa. Leila Denise Alves Ferreira Amorim (Doutora)

Prof. Antônio Eduardo Gomes (Doutor)

Esse trabalho é dedicado a todas e todos que trabalham pela educação no Brasil.

Acknowledgements

Agradeço à minha família. Principalmente à minha mãe, a pessoa mais importante da minha vida. Quem sempre colocou a mim e aos meus irmãos acima de tudo, inclusive dela mesma. Sempre me mostrou o caminho do estudo, sempre me incentivou e abriu todas as portas que podia. Tenho certeza que não estaria aqui se não fosse você. Agradeço aos meus irmãos Filipe, Mateus e Tiago - cada um de vocês me preenche de maneira única. Agradeço à luzinha da minha vida, a minha avó Diva. Também são muito importantes e agradeço muito à minha tia Cássia, meu pai Geraldo e meu avô Nelcílio.

Sou extremamente grata ao Fábio Demarqui e ao Vinícius Mayrink pela orientação, pela paciência, atenção e por estarem sempre disponíveis. Ao Fábio, eu agradeço a empolgação e por me proporcionar a oportunidade de obter todo o conhecimento desses temas tão interessantes. Ao Vinícius eu agradeço por sua calma e dedicação.

Agradeço também às pessoas que colocam essa universidade de pé. Embora às vezes invisibilizadas, elas são fundamentais para que tenhamos tudo em ordem e consigamos trabalhar.

À Juliana Vieira, Jordana Conte e Keilane Pereira eu agradeço. À Ju pelos quinze anos de amizade. Uma das melhores partes de ir ao RJ é encontrar com você e nossas conversas com assuntos que não acabam nunca. À Jordana por sempre se preocupar comigo, por me apoiar, me incentivar e sempre procurar saber como eu estou. À Kei eu agradeço por estarmos sempre em contato, nos apoiando. Também agradeço à amiga Isabella Oliveira.

Pela Liga Externa - Caroline Ponce, Luiz Fernando e Rafael Erbisti - eu agradeço também. As terças que acabavam em pizza viraram encontros com qualquer comida e muita conversa e risada. Vocês são sensacionais. Nossos encontros são sempre maravilhosos.

Sou grata aos presentes que a UFMG me deu: Ana Gabriela, Bárbara da Costa, Cristiano de Carvalho, Douglas Mesquita, Emilly Malveira, Estevão Prado, Fernanda Gabriely, Gabriela Oliveira, Guilherme Oliveira, Guilherme Veloso, Jussiane Gonçalves, Magno Tairone, Rumenick Pereira, Silvana Scheiner e Uriel.

Especialmente, eu agradeço à Bárbara por sua meiguice e doçura. Obrigada trazer luz ao meu dia com a sua chegada, você é muito especial. À Emilly, quem eu considero uma irmã, eu agradeço. Pela companhia, risadas, paciência, confidências. Pela força, pelo suporte e por acreditar em mim. À Gabi, por sempre me fazer pensar “gente, a Gabi é sensacional”. É sempre bom tê-la por perto. Sua animação, conversas, risadas, companhias sempre são um privilégio. À professora Ju por ser sempre uma pessoa doce e solícita. À

Silvana por ter um coração bom e estar sempre disposta a contribuir com as pessoas ao redor.

Ao Estevão eu agradeço por tudo o que você representa na minha vida. Sempre me incentivando e acreditando em mim. Mesmo com as nossas visões de mundo sendo muito diferentes.

Automaticamente eu agradeço à Ana Cláudia, Guilherme Oliveira, Guilherme Veloso e Leticia Nunes. Pela companhia, conversas, risadas e saídas. Vocês são muito especiais pra mim.

Também agradeço ao Otávio Augusto, uma das pessoas com quem eu mais gosto de conversar sobre as coisas do mundo. Sempre aprendo alguma coisa. Ao Ricardo Pedroso pela companhia e por conseguir me aturar.

Às companheiras de casa Évelyn Pôssa e Natália Moller. Fico muito feliz de ter compartilhado parte da minha vida com vocês. Carrego uma parte muito grande de vocês comigo.

I am extremely thankful to Sujit K. Ghosh. Our meetings were always very productive to me. You were always very kind, calm and I learned a lot from you. You inspire me with your intelligence, politeness and dedication. I also thank to Swagata for the lovely and kind person you are.

Sou muito grata à Fernanda Salvato, foi muito incrível ter a sua companhia. Sempre muito animada, solícita e preocupada com o bem-estar das pessoas. Ao Guilherme Pereira pela sempre agradável companhia, pelas conversas e todas as curiosidades que só você sabe. Ao Leonardo Cella, por ser o amor em forma de pessoa. Sua ajuda foi fundamental na minha chegada aos EUA. Agradeço à Juliana Pin e ao Rafael Silva, a companhia de vocês e as conversas foram sempre muito agradáveis. I am also thankful for Dana Johnson, for the kindness and for being the person that I loved to see the most at SAS Hall.

I thank Hojung/Lucy Chan, Gyungyui Lee and Heuiju Chan. To Hojung I thank for the company, for being this sweet person and such a dear friend to me. To Hojung and to Gyungyui I thank for all yoga classes together. To Heuiju I am thankful for the conversation in our office. I also thank SukhDev Mishra for being my friend almost immediately. Always very talkative and teaching me about the fascinating Indian culture and eating habits.

Também sou grata ao professor Luiz Henrique Duczmal pelo excelente curso de Análise. Ao Fredy Walther Castellares Cáceres pela didática ao explicar conceitos de Probabilidade. Ao Wagner Barreto de Souza pela excelente capacidade de ensino.

Agradeço à professora Leila Denise Alves Ferreira Amorim e aos professores Antônio Eduardo Gomes, Dani Gamerman e Enrico Antônio Colosimo por terem aceitado o convite de participar da banca. Ao Enrico eu agradeço novamente por ter acompanhando o

andamento dessa tese praticamente do início ao fim.

Sou extremamente grata ao Serviço de Psicologia Aplicada da UFMG. Em especial ao Alexandre, Daniel, Pedro e Bárbara. Vocês me ajudaram a manter a minha sanidade e a me conhecer. Certamente as sessões mudaram minha vida pra melhor.

Agradeço também à CAPES pela bolsa e pela oportunidade do período sanduíche. Certamente não teria conseguido cursar o doutorado sem a bolsa. Também não haveria a experiência no exterior, que trouxe ótimas contribuições para esse trabalho e para a minha vida pessoal.

Por fim, agradeço de forma geral à todas as pessoas que passaram pela minha vida, me ensinaram alguma coisa, me incentivaram e torceram por mim e que, de alguma forma, contribuíram pra que eu chegasse até aqui.

"Run to the rescue with love and peace will follow".
(River Phoenix)

Abstract

The essence of present work is composed of two main topics, namely: (i) to jointly model longitudinal and survival data; and (ii) to use Bernstein Polynomials (BP) to approximate important and unknown functions in this framework. Considering the joint models, we have, essentially, two main variables: survival times and a variable that is repeatedly measured over time - this latter being called longitudinal variable. We expect that these two variables are related. The structure of the joint model is composed of two sub-models (one for each response variable) that are somehow linked. This type of modeling approach have been used for presenting more precise estimates, since it uses all data information altogether. In turn, Bernstein Polynomials are a very flexible approach and they are used to approximate continuous and smooth functions. Then, our proposal consists of using the BP to approximate the baseline hazard function / cumulative baseline hazard function, as well as the time-varying part of the longitudinal variable. In addition to that, we came up with a solution to the challenge of the choice of the degree. In theory, the larger the degree the better is the approximation. However, in practice, we should consider other aspects, such as the estimation procedure and the concept of parsimony. Thus, the optimal value would be the minimum degree that approximates the main characteristics of the target function. We derived a probabilistic method that indicates a minimum degree, and we also proposed two criteria for the choice of the optimal value. In order to show the benefits of our propositions, we discuss results of two simulation studies. In the first one, we focused on the degree selection method. Then, in the second one, we verified the good performance of joint modeling via BP.

Keywords: baseline hazard function, degree selection, MCMC, `Stan`, time-dependent variable.

Resumo

O trabalho proposto se baseia em dois pontos principais: (i) modelar conjuntamente dados longitudinais e de sobrevivência; e (ii) utilizar os Polinômios de Bernstein (BP) para aproximar funções desconhecidas e de interesse. No contexto de modelagem conjunta, considera-se essencialmente duas variáveis: tempos de sobrevivência e uma variável que é medida repetidas vezes ao longo do tempo - esta última sendo chamada de variável longitudinal. Assim, supõe-se que essas duas informações são relacionadas, e a estrutura dessa modelagem se dá por dois submodelos (um para cada variável) que são ligados de alguma forma. A modelagem conjunta de dados longitudinais e de sobrevivência vem sendo utilizada por apresentar melhorias na estimação, uma vez que explora toda a informação disponível simultaneamente. Por sua vez, os Polinômios de Bernstein se destacam por serem bastante flexíveis, podendo ser uma boa aproximação para qualquer função suave. Nossa proposta é modelar as funções taxa de falha basal / taxa de falha basal acumulada e o comportamento temporal da variável longitudinal através dessa ferramenta. Ainda nesse contexto, foram desenvolvidos critérios para contornar um desafio que consiste na escolha do grau desse polinômio. De forma teórica, quanto maior for o grau, melhor será a aproximação. No entanto, levando em consideração o conceito de parcimônia e a estimação, deseja-se obter um grau mínimo que contemple as características principais da função alvo. Propusemos formas de escolher um grau mínimo a partir de informações *a priori* do comportamento da função de interesse, assim como um grau máximo. Para mostrar os benefícios de utilizar os métodos propostos, foram feitos dois estudos de simulação com réplicas Monte Carlo. O primeiro foca em mostrar o bom funcionamento dos métodos de escolha do grau. Por sua vez, o objetivo do segundo estudo foi verificar a performance da modelagem conjunta via BP.

Palavras-chave: MCMC, seleção do grau, `Stan`, taxa de falha basal, variável tempo-dependente.

List of Figures

List of Tables

Contents

1 Introduction

Within Statistics, it is common to come across data in the form of time until the occurrence of a certain event of interest. In such cases, it is appropriate to make use of methodologies in the field of Survival Analysis. These data are usually attached to some specific characteristics such as the presence of asymmetry and incomplete information, also known as censoring. This incomplete information arises when the event that is being evaluated is not in fact observed (Klein and Moeschberger, 2003). For example, if the event of interest is the failure of a device, it is possible that some of them do not present this failure throughout the entire follow-up period. Such observed times are called "censored times"; whereas those fully observed, that is, the times of those cases in which the failure indeed happened, are called "failure times". Although part of the information collected during follow-up is incomplete, these partial times that were observed (as well as the fully observed times) carry valuable information. Therefore, these observed times are still relevant and they remain on study with the necessity of being treated in a special and adequate manner.

The present study is focused on the most common type of censoring, known as right censoring. Here, the observed time is necessarily lower or equal to the failure time. Another important aspect is the mechanism that causes the failure, which will be considered in this work as independent of that of the censoring.

Since there is a follow-up time for each of the sample elements, there may also be information coming from a longitudinal variable. According to Fitzmaurice et al. (2012), an important reference with regard to longitudinal data and their modeling aspects, these data are characterized by the presence of repeated measurements through time. More specifically, a longitudinal variable is repeatedly measured in the course of follow-up; so, for each sample element, there is one or more observations for it. Then, it is evident that the values of this variable changes with time. Hence, this variable is time-dependent (Faucett and Thomas, 1996) and should be treated as such. One advantage of studies based on longitudinal variables is the possibility to evaluate changes with time of the phenomena being targeted (Xu and Zeger, 2001). This knowledge may be extremely relevant and useful.

As seen in Fitzmaurice et al. (2012), one of the main aspects of longitudinal data is that, regarding the sample as a whole, it is assumed independence *between subjects*, since it is reasonable to consider that the behavior of one individual does not influence the behavior of the others. On the other hand, it is expected to exist considerable correlation between the repeated measurements of the same individual. According to the same authors

mentioned above, these particularities should not be ignored; thus special treatment is necessary.

One of the key topics of this thesis is to consider, and to treat in an adequate manner, cases involving the two types of data. That is, data sets in which there are, essentially, information as: time-to-event, presence of a right censoring scheme, and at least one variable composed by repeated measurements over time. The latter possibly being related to the time to the event.

It is noteworthy that the longitudinal variable is commonly measured with substantial error (Faucett and Thomas, 1996; Wulfsohn and Tsiatis, 1997; Brown and Ibrahim, 2003; Ibrahim et al., 2010). By using these distorted measurements, it is evident that estimates related to the risk of the occurrence of the event under study are obtained with bias. In addition, when there is a long follow-up period, the longitudinal variable is also subject to missing data. And the reason for this missingness may be related to the study itself.

At first, in order to model data containing longitudinal and survival variables, one may consider one of these elements (Wu and Bailey, 1988). As an example, survival analysis could be applied with an adaptation, aiming to include in some way the repeated measurements of the longitudinal variable (Wu et al., 2012). This path goes against the extremely reasonable idea that the occurrence of the event of interest implies on changes in the values of the longitudinal variable and vice versa; that is, it does not take into account the mutual relationship of dependence between these two variables. One consequence of using this approach is the presence of biased estimates (Wu et al., 2012).

Joint modeling of longitudinal and survival data comes up with improvements in the sense of more accurate estimates. Furthermore, as stated on Tsiatis and Davidian (2004) and Wu et al. (2012), it also provides important information such as how the behavior of the longitudinal variable evolves with time, the risk of the occurrence of an event, and the relationship between the longitudinal variable and the survival of patients. Another advantage of this model is the possibility of contemplating, in a relatively simple way, the matter involving measurement error. The gain here relies on the perception that using "true" values, that is, estimates representing the actual true values, are more appropriated comparing to the biased (and observed) ones (Brown and Ibrahim, 2003).

According to Ibrahim et al. (2010), initial works addressing joint modeling of longitudinal and survival data were motivated by Human Immunodeficiency Viruses (HIV) studies. Under this scenario, the event of interest often was the disease progression or death, and the longitudinal variable was the CD4 cell count. This variable is known to be a marker for the disease progression.

Faucett and Thomas (1996) used the Bayesian approach to model survival data

with a continuous longitudinal variable prone to measurement error. A sub-model was established to describe the process of the longitudinal variable and another one concerning the survival part. These two sub-models were linked through a parameter on the survival sub-model, and all parameters were estimated simultaneously. The results discussed by the authors indicated that better estimates were obtained, especially referring to the parameters related to the the survival component.

Many other works, such as Wulfsohn and Tsiatis (1997), Xu and Zeger (2001), Brown and Ibrahim (2003) and Ibrahim et al. (2010) applied this same structure of joint modeling. Comparing with the work of Faucett and Thomas (1996), there exists improvements, since there was no restriction on the type of the longitudinal variable. The first and last studies made use of frequentist inference procedures, while the others were based on the Bayesian approach. More recently, Brilleman et al. (2017) described how to implement joint models for longitudinal and survival data in the `Stan` platform (Carpenter et al., 2017). This paper discusses both theoretical and computational aspects of this model.

In what follows, the second main topic of the present work is related to the techniques employed to the estimation of some of the unknown functions of interest. This estimation will be done via Bernstein Polynomials (BP) (Bernstein, 1912). The BP were developed by Sergei Natanovich Bernstein in 1912 when he was proving a demonstration to a special case of Weiertrass' Theorem (Bernstein, 1912; Lorentz, 1986). The usage of polynomials to approximate functions has analytic advantages since they are easily written in the form of summation. As a result, and highlighted by Osman and Ghosh (2012), calculations such as derivatives and gradients are easier to obtain through this structure. The main utility of the BP lies on approximating any smooth curve/function. An additional reference is Kuller (1964), who replicated the original paper, focusing on giving details about the demonstration.

In the statistical literature, authors have been using BP in several and diverse situations: Vitale (1975), Petrone (1999a), Petrone (1999b) and Babu et al. (2002) make use of BP in the context of density estimation. Kottas (2006) applied the BP for density as well as intensity estimation function in a Poisson Process model. Brown and Chen (1999) use the BP as kernel estimator. Moreover, Bernstein Polynomials can be used aiming at variable selection, as in Curtis and Ghosh (2011), and in shape-restriction problems like in Chang et al. (2007) and Wang and Ghosh (2012).

Chang et al. (2007) and Wang and Ghosh (2012) used BP to approximate shape restricted curves. These characteristics are relatively easy to obtain when using the BP, by simply imposing conditions on a vector that composes this polynomial. They discussed monotone, convex/concave restrictions, among others. This particularity of BP manages to guarantee that the estimated curve/function maintains essential characteristics of the

target function. Chang et al. (2007) also discussed the relationship between a part that composes the BP and the number of roots in the curve being approximated.

Curtis and Ghosh (2011) made use of BP in the case of variable selection. The motivation lied on possible non-linear relationships between a response variable and continuous covariates. In this case, the role of BP was to approximate a curve representing this relationship. The benefit of this strategy is that there is no need to previously impose a structure (*i. e.* linear, quadratic, etc.). It can be considered unknown and BP will be able to approximate it.

Farouki (2012) provides an extensive reference about Bernstein Polynomials and the Bernstein basis. This reference brings attention to the personal history of Sergei Bernstein - the person who came up with the BP -, its mathematical properties, historical aspects, relationship with the monomial and other bases - such as splines, Bézier and de Casteljau. Extra discussions in this paper focus on specific details about BP performance, numerical stability, convergence, advantages and disadvantages.

More specifically to the focus of this thesis, which is longitudinal and survival analysis, we were not able to find many works. Chang et al. (2005) used Bernstein Polynomials to approximate the cumulative hazard function considering right censorship scheme; Osman and Ghosh (2012) used BP for approximating the baseline cumulative and hazard functions considering non-proportional hazards. An interesting point of the first work is the contemplation of the meaning of the coefficients of the BP; however, they did not include the information from covariates in the modeling procedure. Chen et al. (2014) used the BP to approximate baseline survival functions under the accelerated hazards model for right-censored subjects. They also considered time-dependent covariates.

Zhou et al. (2017) considered interval-censored bivariate survival data. In this paper, the BP was applied to approximate the baseline hazard function under a frequentist approach. Their proposed model includes as special cases, the proportional hazard model and the proportional odds model. An interesting point in their work is that they claimed that their method seemed to be robust concerning the degree of the BP; but, perhaps, this result may be more related to the simplicity and smoothness of the baseline hazard function than associated to the robustness of the method itself. In a somewhat similar scenario, Zhou and Hanson (2018) treated interval-censored spatial-referenced survival data using BP to model the baseline survival function. The proposed approach includes the proportional hazards model, the proportional odds model and accelerated failure time model.

Bertrand et al. (2019) emphasizes the importance of considering properly the information from covariates measured with error. This relevance motivated them to use the BP to model covariates with this type of peculiarity. They applied the proposed method in a context of survival analysis in a cross-section simulated data.

An important discussion is how to choose an optimal degree for the BP. This degree is strictly related to how close the estimated curve will be from the true one. If the degree is too small, the approximation may not be flexible enough to include changes in the curve being approximated, for example. However, if this quantity is too large there may be computational issues since it implies in an increasing number of parameters. In order to solve this issue about the degree, some works like Curtis and Ghosh (2011), Osman and Ghosh (2012), Wang and Ghosh (2012), Zhou et al. (2017), Zhou and Hanson (2018), and Bertrand et al. (2019) fixed this number and several of these works discussed strategies on how to choose it properly. Nonetheless, Petrone (1999a), Petrone (1999b), Chang et al. (2005), Chang et al. (2007) and Chen et al. (2014) consider the degree of the BP as random quantity to be estimated. In this thesis, we studied the probability of a minimum suitable degree as well as two criteria that serve as a stopping rule to establish a maximum degree.

To the best of our knowledge, only Guan (2016) proposed an estimator for the degree of the BP and studied its performance. The mentioned proposal is different from ours. Our method seems to be more attractive in the sense that it is based on probabilities and an entire distribution can be studied. Also, Guan's method is based on the frequentist approach. On the other hand, the application of our method requires a previous knowledge of where the target function will change behavior to establish a minimum degree, and a posterior sample of the vector of coefficients to choose an optimal degree.

The main contributions of this thesis is the joint modeling of longitudinal and survival data via Bernstein Polynomials. Here, we used the structure of the BP to model both variables. In addition to that, we propose a solution to the challenge of choosing a degree for the BP. We also showed via simulation study the good performance of the BP and of our degree selection method.

This thesis is organized as follows: on Chapter 2 we introduce some basic concepts about longitudinal and survival data. In the mentioned chapter, we also discuss separate and joint ways of modeling these variables. Our aim was to improve the understating of important topics of this thesis. In turn, Chapter 3 consider the one of the main topics of this study, which is the Bernstein Polynomials. Then, in Chapter 4, we show and debate results of simulation studies. Chapter 5 contemplates an application of real data illustrating the results obtained so far. At last, Chapter 6 shows some of the next steps as well as ideas for future works.

2 Basic Concepts

One of the main goals of the present work concerns on properly handling longitudinal and survival data. Then, the focus of this chapter is to introduce basic concepts related to each of these types of data. In Section 2.1, we describe longitudinal data and discuss some of its characteristics and properties. We also mention a usual modeling procedure. Next, in Section 2.2, we explain basic topics of survival data analysis. Some of these topics are the description of survival functions, the peculiarities and challenges involving these data, and modeling approaches. The brief discussion we present in these two sections will enhance a better understanding of the joint model framework, which is defined and described in Section 2.3.

2.1 Longitudinal Data

Longitudinal data are fundamentally characterized by repeated measurements of one or more variables for each specific subject of a sample (Fitzmaurice et al., 2012). For instance, we can think of a study involving seropositive patients. It is widely reported in literature that the CD4 cell count is a measure of progression of this disease. For this reason, this amount is usually accompanied for each patient throughout the study. The implication of this procedure is that there will be more than one observation (repeated measurement) of CD4 cell count for each patient. Then, considering this characteristic that defines longitudinal data, it is clear that this variable changes with time and, therefore, it is time-dependent.

One of the main aspects of longitudinal data is that, regarding the sample as a whole, it is reasonable to assume independence *between individuals*. This feature is a regular assumption in other studies and methods. However, it is expected to exist significant correlation *inner individual*, precisely because longitudinal variables are measured multiple times for the same subject. As a result, the usual methods of analysis such as linear or generalized linear models for example, may not be suitable here.

According to Fitzmaurice et al. (2012), the correlation between the repeated measurements is usually positive. For example, we can think of a study with hypertensive patients, and the longitudinal variable being the blood pressure. It is expected that, if the subject's blood pressure is high, it will continue to be high in the long run even if a treatment manages to reduce it now. Besides that, the same authors affirm that this correlation between measurements rarely gets close to 1, even when comparing measurements within tiny gaps of time.

Another characteristic that is commonly found in longitudinal studies is missing data. These missing information occur because, usually, such studies require a long period of follow-up. Then, it often happens that a patient misses an appointment or, perhaps, she/he is too ill to have her/his measurement collected. Therefore, these situations "generate" missing values on the data set. To simply discard or ignore these missing observations may lead to bias on the analyses (Tsiatis et al., 1995), and this bias can be more troublesome if the reason that causes the missingness is related to the problem being studied itself.

The presence of longitudinal components on data set enriches the analysis by allowing specific elucidations and knowledge about subjects under study. It is possible, for example, to compare individuals with her/his repeated measurements over time. So, there is homogeneity in comparisons, since we are contrasting the same subject and her/his respective evolution in time. For this reason, it is conceivable to evaluate the evolution of the longitudinal variable at individual level or, in other words, how it changes through time for every single subject (Fitzmaurice et al., 2012), detecting highs, lows and even trends.

It is worth pointing out the difference between balanced and unbalanced data sets. As it was affirmed previously, longitudinal measurements are taken at some points of time. These times can be represented by t_{ij}, for subjects $i = 1, 2, \ldots, n$ and for time points $j = 1, 2, \ldots, J_i$. Note that the subjects under study may have different number of measurements J_i, as well as measurements registered at different times. Given this, if all J_i and t_{ij} are equal for all subjects, the data set is said balanced; otherwise it is unbalanced. In addition, an unbalance data set may be due to the number of measurements and/or to the times in which the measurements were evaluated.

A final comment about this type of data is that they are usually prone to measurement error (Ibrahim et al., 2001; Klein et al., 2013). It is very intuitive to perceive - and there are several studies that endorse this idea - that it is better to model this variable in order to have an approximation of what the true value is, than to just use raw data. A framework that fits this idea is the Mixed Effects (ME) Model, for example. It will be briefly described in the next section.

2.1.1 Mixed Effects Models

We can use ME Models as an alternative to handle longitudinal variables (Harville, 1977; Laird and Ware, 1982). The structure of these models includes both fixed and random effects. Thus, for subject $i = 1, 2, \ldots, n$ and a time point $j = 1, 2, \ldots, J_i$, we can write the linear ME model, in the longitudinal analysis context, as follows:

$$\begin{aligned} Y_i(t_{ij}) &= W_i(t_{ij}) + \epsilon_i(t_{ij}) \\ &= \mathbf{x}_i\boldsymbol{\beta} + \mathbf{b}_i^\top \mathbf{f}(t_{ij}) + \epsilon_i(t_{ij}), \end{aligned} \tag{2.1}$$

where $Y_i(t_{ij})$ is the random variable that represents the observed value of the longitudinal variable for the i-th subject at time t_{ij}. Assuming that it was measured with error, $W_i(t_{ij})$ is the true and unobserved value of this variable. In turn, $\epsilon_i(t_{ij})$ is the measurement error following a $Normal(0, \sigma_\epsilon^2)$ distribution. The vector of regression coefficients is composed of $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$, and the vector of covariates associated with these coefficients is $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top$. Finally, the vector of random effects at subject level is represented by $\mathbf{b}_i$. This vector is associated with a vector of functions of time $\mathbf{f}(.)$, and it does not depend on ϵ_i. The distribution of $\mathbf{b}_i$ is usually $Normal_p(\boldsymbol{\mu}_b, \Sigma_b)$, for $i = 1, 2, \ldots, n$. We highlight here that a correlation structure may be included via the variance-covariance matrix Σ_b.

The interpretation of the fixed effects concerns the sample as a whole, *i. e.*, at population level. Then, the vector of coefficients $\boldsymbol{\beta}$ tells the impact of the vector of covariates in the mean of the longitudinal variable. In addition to that, the vector of overall means $\boldsymbol{\mu}_b$ describes the true characteristics of the behavior of this variable and its variation along the time. Then, each subject has its own variation over the overall mean. This variation is expressed through the vector of random effects $\mathbf{b}_i$, for $i = 1, 2, \ldots, n$. With this vector in hand, we can also verify the heterogeneity between subjects, for example.

When we assume a Normal distribution for the measurement error $\epsilon_i(t_{ij})$, it is immediately implied that the conditional distribution $Y_i(t_{ij})|\mathbf{b}_i$ is normally distributed with parameters $\mathbb{E}\left[Y_i(t_{ij})|\mathbf{b}_i\right] = W_i(t_{ij}) = \mathbf{x}_i\boldsymbol{\beta} + \mathbf{b}_i^\top \mathbf{f}(t_{ij})$ and $\mathbb{V}ar\left[Y_i(t_{ij})|\mathbf{b}_i\right] = \sigma_\epsilon^2$. This result has an interesting meaning that, given the vector of random effects $\mathbf{b}_i$, the expected value of the response variable is the "true" value and it varies according to the variance of the measurement error. On the other hand, the marginal distribution of $Y_i(t_{ij})$ is also Normal, with mean $\mathbb{E}\left[Y_i(t_{ij})\right] = \mathbf{x}_i\boldsymbol{\beta} + \mathbf{f}(t_{ij})\boldsymbol{\mu}_b$ and variance equal to $\mathbb{V}ar\left[Y_i(t_{ij})\right] = \sigma_\epsilon^2 + \left[\mathbf{f}(t_{ij})\right]\Sigma_b\left[\mathbf{f}(t_{ij})\right]^\top$. The calculations of the marginal distribution of this random variable are shown in the Appendix A, page 99.

Considering what was exposed above, the likelihood function is given by

$$\begin{aligned} \mathcal{L}(\boldsymbol{\Phi}; Data) &= \prod_{i=1}^{n} p(\mathbf{y}_i|\boldsymbol{\Phi}) = \prod_{i=1}^{n} \int p(\mathbf{y}_i, \mathbf{b}_i|\boldsymbol{\Phi}) d\mathbf{b}_i = \prod_{i=1}^{n} \int p(\mathbf{y}_i|\mathbf{b}_i, \boldsymbol{\Phi}) p(\mathbf{b}_i|\boldsymbol{\mu}_b, \Sigma_b) d\mathbf{b}_i \\ &= \prod_{i=1}^{n} \int \left\{ \prod_{j=1}^{J_i} p(y_i(t_{ij})|\mathbf{b}_i, \boldsymbol{\Phi}) p(\mathbf{b}_i|\boldsymbol{\mu}_b, \Sigma_b) \right\} d\mathbf{b}_i \\ &= \prod_{i=1}^{n} \int \frac{1}{(2\pi\sigma_\epsilon^2)^{J_i/2}} \exp\left\{ -\sum_{j=1}^{J_i} \frac{\{y_i(t_{ij}) - W_i(t_{ij})\}^2}{2\sigma_\epsilon^2} \right\} p(\mathbf{b}_i|\boldsymbol{\mu}_b, \Sigma_b) d\mathbf{b}_i, \end{aligned} \tag{2.2}$$

where $\boldsymbol{\Phi} = \{\boldsymbol{\beta}, \sigma^2_\epsilon, \boldsymbol{\mu}_b, \Sigma_b\}$ represents the set of unknown parameters. The entire data set is expressed by $Data = \{\mathbf{t}_i, \mathbf{y}_i, \mathbf{x}_i,\ i = 1, 2, \ldots, n\}$; such that, for the i-th subject, $\mathbf{t}_i = (t_{i1}, t_{i2}, \ldots, t_{iJ_i})$ is the vector of measurement times, $\mathbf{y}_i = (y_i(t_{i1}), y_i(t_{i2}), \ldots, y_i(t_{iJ_i}))$ is the vector of observed values of the longitudinal variable, and $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is the vector of covariates. The sample size is represented by n and J_i is the number of measurements for each subject i. Therefore, the total number of observations is given by $\sum_{i=1}^{n} J_i$. The vector of random effects at subject level is $\mathbf{b}_i$, and we will consider this vector following a Normal distribution with mean $\boldsymbol{\mu}_b$ and variance-covariance matrix Σ_b. At last, $\mathbf{W}_i = (W_i(t_{i1}), W_i(t_{i2}), \ldots, W_i(t_{iJ_i}))$ is the vector of the non-observed true values of the longitudinal variable.

We will use the structure of the ME models described above on the longitudinal variable for joint modeling longitudinal and survival data.

2.2 Survival Data

This section addresses few basic and important concepts in the field of survival analysis. The functions described here are essential, since they will be constantly considered throughout this thesis.

Survival analysis is the area of statistics that contemplates response variables in the form of "time until the occurrence of an event of interest". The commonly used way to express the occurrence of this event is "failure", regardless of what is this event being studied. Some examples are the time until smoke cessation, death, recurrence of a disease, failure of a machine, customers delinquency, among others. Some implications related to this type of data is the presence of asymmetry and partially observed times, *i. e.*, incomplete information. These incomplete information arises when the event under study is not indeed observed. In this case, the observed time is called a "censored time"; whereas "failure times" are those in which the time until failure actually was accounted.

There are three types of censoring: left, interval and right censoring (Klein and Moeschberger, 2003). In the present study we will consider the situation of a right censoring scheme. This case is the one that happens the most in practice. Under this scenario, if a subject is censored, the true time to failure is unknown, but it is known that it is greater than the observed time. Some examples of situations that imply on right censoring are: study dropout, loss of follow-up and death by other causes than the one being studied. We define an indicator variable δ_i to represent whether the i-th subject has experienced the event under study, or if it was right censored. Then, for subjects $i = 1, 2, \ldots, n$ we have that:

$$\delta_i = \begin{cases} 1, \text{ if subject } i \text{ has failed;} \\ 0, \text{ if subject } i \text{ is (right) censored.} \end{cases}$$

In addition, consider T_i a continuous random variable representing the time to failure and C_i the time until censoring, both for subject $i = 1, 2, \ldots, n$. The observed time u_i is either a failure ($\delta_i = 1$) or a censoring ($\delta_i = 0$), whichever happened first; that is, $U_i = \min(T_i, C_i)$. Thus, for every subject under follow-up, the observed data are of the form (u_i, δ_i).

Some important functions are the basis for much of the inferential processes in survival analysis. They are: the survival function, the hazard function and the cumulative hazard function. The survival function $S(.)$ is defined as the probability that the time to event is greater than a value $u > 0$. That is,

$$S(u) = \mathbb{P}(T > u) = 1 - \mathbb{P}(T \leq u) = 1 - F(u),$$

where $F(.)$ is a cumulative distribution function. The survival function has three important properties, they are

1. $S(0) = 1$;
2. $\lim_{u \to \infty} S(u) = 0$;
3. $S(u)$ is a non-increasing function of time u.

The first property means that at the beginning of follow-up all subjects are "alive", that is, they did not suffered the event of interest yet. The second one implies that if the follow-up time is long enough ($u \to \infty$), all subjects will suffer the event under study. It also means that there is no cure fraction; see more about this topic in Ibrahim et al. (2001), Rodrigues et al. (2008) and Klein et al. (2013). At last, the third property is inherited directly from the relationship with the cumulative distribution function $F(.)$.

The hazard function, denoted by $h(.)$, is the instantaneous failure rate. It can be understood as the risk of occurrence of the event under study. Its expression is given by

$$h(u) = \lim_{du \to 0} \frac{\mathbb{P}(u \leq T < u + du | T \geq u)}{du}.$$

Finally, the cumulative hazard function refers to the risk of failure *until* time u:

$$H(u) = \int_0^u h(t)dt.$$

It is important to mention that these three functions are related to one another. The implication of this relationship is that, by specifying a single one of them, all the others are automatically defined. The main relationships are:

$$S(u) = \exp(-H(u)) \qquad \text{and} \qquad h(u) = \frac{f(u)}{S(u)}.$$

Based on the expressions defined above, we can write the likelihood function. In the present study, as stated previously, we will consider a right censoring scheme and a non-informative censoring mechanism. This non-informative mechanism means that the time to failure and time to censorship are considered independent. Therefore, the individual contribution to the likelihood function will be the density function in case of failure, and the survival function for the right censored cases (Lawless, 2003). That is,

$$\begin{aligned} \mathcal{L}(\boldsymbol{\Phi}; Data) &\propto \prod_{i=1}^{n} f(u_i|\boldsymbol{\Phi})^{\delta_i} (S(u_i|\boldsymbol{\Phi}))^{1-\delta_i} \\ &\propto \prod_{i=1}^{n} h(u_i|\boldsymbol{\Phi})^{\delta_i} \exp\left\{-H(u_i|\boldsymbol{\Phi})\right\}, \end{aligned} \tag{2.3}$$

where $\boldsymbol{\Phi}$ is the vector of all parameters to be estimated, and $Data$ represents all data available. In this case, $Data = \{u_i, \delta_i, \text{ for } i = 1, 2, \ldots, n\}$, in which u_i is the observed survival time, δ_i is the failure/censoring indicator for subject i and n is the sample size.

Generally speaking, one of the goals when analyzing survival data is to obtain good approximations for the survival and/or hazard function. Thus, it is desired to make use of flexible structures to model these functions. In order to do so, there are non-parametric, and parametric methods available. In addition, in the case of the presence of covariates, we can also consider semi-parametric methods.

The most popular non-parametric estimator is the Kaplan-Meier (KM) (Kaplan and Meier, 1958). This option provides an approximation for the survival function based on distinct observed failure times. It can also be used to analyze data descriptively.

Taking into consideration parametric methods, we assume a probability distribution for the time to failure T_i, and then we obtain the survival functions. After that, once the likelihood function in Equation (2.3) is completely specified, we make use of inferential methods to obtain parameter estimates. A potential drawback in such approach is that it can be restrictive, in the sense that the possible shapes of the hazard function are already pre-specified. Nonetheless, there are very flexible distributions, like the Piecewise Exponential (PE) (Arjas and Gasbarra, 1994) and the Birnbaum-Saunders (Birnbaum and Saunders, 1969), for example.

Finally, as an example of a popular semi-parametric method we can cite the proportional hazards model (Cox, 1972). A more detailed description of this alternative is given in the next section.

2.2.1 Proportional Hazard Model

A widely used model, in the context of survival analysis, is the proportional hazards model (Cox, 1972). This is a regression model and it is defined via the hazard function. The structure is given as follows:

$$h(u|\mathbf{z}) = h_0(u) \exp\left\{\mathbf{z}^\top \boldsymbol{\psi}\right\}, \text{ for } u > 0, \tag{2.4}$$

where $h_0(u)$ is the baseline hazard function at time u and $\boldsymbol{\psi}$ is a vector of coefficients associated with the vector of covariates $\mathbf{z}$. The interpretation of the baseline hazard function can be related to that of an intercept. That is, how the hazard function behaves when all the covariates are equal to zero, for continuous covariates; or belong to the reference group, in the case of categorical ones.

One of the most important assumptions related to this model is that the ratio of the hazards between two subjects is proportional in time. In order to show this mathematically, we consider $\mathbf{z}_1$ as the vector of covariates for one subject, and $\mathbf{z}_2$ for another subject. Then, the hazard ratio is given by

$$\frac{h(u|\mathbf{z}_1)}{h(u|\mathbf{z}_2)} = \frac{h_0(u) \exp\left\{\mathbf{z}_1^\top \boldsymbol{\psi}\right\}}{h_0(u) \exp\left\{\mathbf{z}_2^\top \boldsymbol{\psi}\right\}} = \frac{\exp\left\{\mathbf{z}_1^\top \boldsymbol{\psi}\right\}}{\exp\left\{\mathbf{z}_2^\top \boldsymbol{\psi}\right\}} = \exp\left\{(\mathbf{z}_1 - \mathbf{z}_2)^\top \boldsymbol{\psi}\right\}. \tag{2.5}$$

Note, in Equation (2.5), that the result does not depend on the time u.

In the original characterization of the proportional hazards model in Equation (2.4), the baseline hazard function $h_0(.)$ is left unspecified (Cox, 1972). It is also possible to model this function by assuming a distribution for the failure times. This alternative leads to the so called parametric Cox model.

Several extensions were proposed to the proportional hazards model. One of them will be quite explored in this work: the proportional hazards model for time-dependent covariates. Time-dependent covariates are those with values not necessarily constant over time. Some examples are: blood pressure, size of tumor, treatment (since it can change due to the efficacy of a treatment or according to the state of the health of the patient), CD4 cell count and many others. The mentioned model is given by:

$$h(u|\mathbf{z}(u)) = h_0(u) \exp\left\{\mathbf{z}(u)^\top \boldsymbol{\psi}\right\}, u > 0, \tag{2.6}$$

here, $\mathbf{z}(u)$ is the vector of time-varying covariates.

A restrictive and, in many cases, unrealistic assumption related to the proportional hazards model for time-dependent variables, is the requirement of knowing the values of $\mathbf{z}(.)$ for all time t that the subject is under observation, *i. e.* $0 < t < u$, with u representing the observed survival time (Klein and Moeschberger, 2003). In the vast majority of practical cases that we are aware of, only a few values are indeed measured. In this model, the proportional hazards property is no longer true due to the presence of a time-dependent variable (Colosimo and Giolo, 2006). In this case, the hazard ratio is

$$\frac{h(u|\mathbf{z}_1(u))}{h(u|\mathbf{z}_2(u))} = \frac{h_0(u)\exp\left\{\mathbf{z}_1(u)^\top\boldsymbol{\psi}\right\}}{h_0(u)\exp\left\{\mathbf{z}_2(u)^\top\boldsymbol{\psi}\right\}} = \frac{\exp\left\{\mathbf{z}_1(u)^\top\boldsymbol{\psi}\right\}}{\exp\left\{\mathbf{z}_2(u)^\top\boldsymbol{\psi}\right\}} = \exp\left\{(\mathbf{z}_1(u)-\mathbf{z}_2(u))^\top\boldsymbol{\psi}\right\},$$

which clearly depends on the time u.

The next section describes how to jointly model longitudinal and survival data. Concepts regarding both of these variables will be resumed there.

2.3 Joint Models for Longitudinal and Survival Data

This section concerns the introduction, explanation and discussion of joint models for longitudinal and survival data. This topic is one of the two main themes that build the essence of this thesis. Some important concepts that have already been introduced will be considered again and contextualized according to the specific purposes of this matter.

In many studies, it is common to record data including a time to event variable as well as longitudinal and, consequently, time-dependent covariates. In cancer studies, for example, there may be survival information such as time until death or time until recurrence. In this case, the time-dependent/longitudinal variables, can be the size of tumor or a variable representing the quality of life (Ibrahim et al., 2010). Another study in which this type of data is very common is with seropositive patients (Ibrahim et al., 2010). Here, the event of interest may be death, or disease progression, and the longitudinal variable that is usually evaluated is the CD4 cell count. The motivation to collect data in this structure is that it is already established in literature that CD4 cell count is a disease progression marker. We will recapitulate this example of seropositive patients ahead in this thesis aiming at a better understanding of the concepts we discuss.

In the given examples there are at least two main variables of interest: the longitudinal and the time to an event. As a result, one possible way to model such data is to prioritize either the longitudinal or the survival variable. Then, if we focus on the longitudinal variable, we can apply the ME model for example (see Section 2.1.1 on page 21). However, we can use other options such as a Non-Linear Mixed Effects Model or a

Generalized Linear Mixed Effects Model. On the other hand, if we are more interested on the outcomes related to the survival variable, we can make use of the proportional hazards model for time-dependent covariates (described in Section 2.2.1, page 26); see also Ibrahim et al. (2010) and Rizopoulos (2012).

In the case of using the proportional hazards model for time-dependent covariates, there can be an inconvenience. This inconvenience is that this model requires the knowledge of the values of the longitudinal variable for every unity of the follow-up time. That is, for all time t, $0 < t < u$, with u being the observed survival time (Tsiatis et al., 1992; Klein and Moeschberger, 2003; Rizopoulos, 2012). Then, if a subject was followed-up for a certain amount of time unities, there should be this same number of measurements for the longitudinal variable. Nevertheless, in many cases, as we have already discussed, these measurements are only collected on specific time points (t_{ij}, for subjects $i = 1, 2, \ldots, n$ and time points $j = 1, 2, \ldots, J_i$). Thus, this is not a realistic requirement if we consider real data.

Time-dependent variables can be classified as internal or external variables. Alternative names for these classifications are endogenous and exogenous, respectively. The first designation refers to the case where the evaluation of this quantity depends on the subject being examined herself/himself. Some examples of internal variables are the size of tumor, CD4 cell count, blood pressure, among others. Therefore, the occurrence of the event of interest may prevent these measurements from being collected. That is, not only the value of the longitudinal variable, but also the number of repeated measurements over time may depend on the survival. This situation may have a direct impact on estimation. In turn, external or exogenous covariates does not depend on the subject. As examples, we can cite temperature and air pollution (Rizopoulos, 2012).

Wu et al. (2012) mention that a separate analysis of longitudinal and survival data may lead to biased estimates for even more reasons. It is natural and intuitive to expect that the occurrence of the event of interest affects the behavior of endogenous longitudinal variables. Besides that, the length of follow-up time might also interfere on this variable. In a separate analysis this information is not fully taken into account (Ibrahim et al., 2010). In addition, since longitudinal variables are generally prone to measurement error and may have missing data (Ibrahim et al., 2010; Wu et al., 2012), it is important to handle all these elements. Simply using raw data in proportional hazards model for time-dependent covariates may produce poor and inconsistent estimates (Wu et al., 2012).

One of the first efforts to address both longitudinal and survival variables in a more robust procedure, was done by Tsiatis et al. (1995) and Faucett and Thomas (1996). Their proposal was called two-stage models. This approach considers the two main variables in the same analysis with a single modeling structure for both of them. However, the estimation procedure is by means of two steps. Thus, their proposal was an improvement when

compared to the separate analysis. Fundamentally, the idea is to model the longitudinal variable on the first stage; and then, in the second stage, we use these treated values to model the survival response. That is:

> **1^{st} stage, longitudinal sub-model**: in this stage we model the longitudinal variable, treating possible measurement errors and missing values when convenient. These estimations represent the unobserved true measurements;
>
> **2^{nd} stage, survival sub-model**: we model survival data conditioned on the estimation results of the first stage.

The two-stage model has advantages and drawbacks. In what it refers to the gains obtained in using this model, we can cite the simplicity and the ease to implement (Wu et al., 2012). This convenience is in relation to both mathematical and computational aspects. On the other hand, the dependence relationship between both longitudinal and survival components is not taken into account at the first stage. In addition to that, the information of the occurrence of the event of interest is not considered in this part as well. The impact of ignoring these information at the first stage is that, in the case of internal longitudinal variables, the occurrence of the event influences how many times we will measure the longitudinal variable, and also the values that we will observe (Wu et al., 2012). Another point to consider as a disadvantage is that, since we use only the point estimates obtained in the first stage to model the survival variable at the second stage, the variability of these estimates are ignored. Moreover, through a simulation study, Wu et al. (2012) verified that this model underestimates parameters that link the two sub-models and its variability. Another interesting result, according to these same authors, is that the bias on the first stage is more related to the strength of the relationship between both processes; while the bias towards the parameters of the survival component depends on the magnitude of measurement error.

In turn, joint models for longitudinal and survival data emerged aiming at improving the negative aspects of both separate and two stage models. The initial practical motivation, according to Ibrahim et al. (2010), was based on HIV studies. Joint modeling provides more robust estimates as it uses all information altogether; see also Wu et al. (2012). In addition, Ibrahim et al. (2010) state that these models reduce bias on parameter estimates. According to Tsiatis and Davidian (2004), Ibrahim et al. (2010), Rizopoulos (2012) and Wu et al. (2012), by using joint models we are capable to obtain interesting and useful information such as:

(i) the trajectory of individuals related to the longitudinal variable. That is, how this variable varies and behaves with time for each subject of the sample. This is very important because it provides knowledge to the researcher about which periods of

time these values will be too high or low, for example, helping the decision making processes;

(ii) survival of patients, as it is usual and very relevant on the framework of survival analysis. By using joint models, we can infer about characteristics that impacts on the occurrence of the event under study with more precise estimates;

(iii) the relationship between longitudinal and survival variables, since we expect that the values and the behavior of the longitudinal variable influences the survival outcome and vice-versa.

Resuming the discussion about internal and external time dependent covariates (see page 28), Rizopoulos (2012) and Wu et al. (2012) affirm that joint models are indicated especially for treating internal longitudinal variables. This indication is based on the idea that the occurrence of the event of interest intervenes directly on the measurements of the longitudinal variable, and the structure of joint models is able to take this information into account.

Next, we will take back the case of an HIV study with the longitudinal variable being the CD4 cell count. We mentioned previously that this quantity is known in the medical literature to be an important surrogate marker to the disease progression. Then, if the count of this variable is low, we can expect that the event will occur more rapidly for that specific individual. On the other hand, if it is relatively high, than the subject might be with a better health status. By using the joint modeling approach, we can assess and comprehend the ups and downs of CD4 cell quantity over time for each subject under study. In addition, these estimates are obtained considering all the information such as the covariates, survival times, failures and censorings, and their relationships. And these information are accounted altogether. Moreover, we have interpretations for the survival aspects. For this other variable we can tell the role of CD4 cell count to explain the time until the appearance of an opportunistic disease or death.

The structure of the joint model for survival and longitudinal data is composed by two sub-models. These sub-models are connected in some way and all parameters are estimated simultaneously (Tsiatis and Davidian, 2004; Ibrahim et al., 2010). A simple and commonly used structure for the longitudinal sub-model is the ME model. In turn, we can use the accelerated failure time model (see Lawless (2003) for more information) or proportional hazards model for time dependent covariates to model the the survival component (Wu et al., 2012). The mentioned link between the two sub-models may be a shared random effect or a coefficient, for example. In the list below, we cite the procedures that one can follow to jointly model these variables. Then, we explain each one of them.

1. Define a function for the longitudinal variable whose behavior is able to properly

represent the evolution over time for this variable. This function is called the trajectory function;

2. specify a modeling structure for the longitudinal component based on the observed values;

3. determine a proper model for the survival component.

We mention, in advance, that there is a connection between these steps. We will explain this connection ahead.

The role of the trajectory function is to represent the values of the longitudinal variable for all time t, such that $0 < t < u$. Therefore, we will have representations for the unobserved true values, *i. e.*, values considered free of measurement error. In addition, we will also have values for time points in which no measurement was taken. In other words, since this function is defined for all time unities, it provides information about the evolutionary aspects of the longitudinal and also time-dependent variable. A possible way to define the trajectory function is the following (Tsiatis and Davidian, 2004):

$$W_i(t) = \mathbf{f}(t)\mathbf{b}_i,\ t > 0,\ i = 1, 2, \ldots, n. \tag{2.7}$$

In Equation (2.7), the term $W_i(t)$ is the trajectory function for the i-th individual on time t. This function represents the true value of the longitudinal variable on that time point, whether we have an observation for this time or not. In turn, $\mathbf{f}$ is a vector of continuous functions and $\mathbf{b}_i$ is a vector of random effects. These random effects may be, for example, a random intercept and slope at individual level. In this case, $\mathbf{f}(t)$ would be the vector $(1, t)$. The formulation in this equation includes the case of linear trajectory, polynomials, splines and other non-linear functions on time t (Tsiatis and Davidian, 2004).

We highlight here that we can have other specifications for the vector of functions $\mathbf{f}$. For example, if we are interested in capturing specific nuances of the true biologic process, we can add a stochastic process in Equation (2.7) (Tsiatis and Davidian, 2004; Rizopoulos, 2012). However, in order to determinate the trajectory function we must consider the trade-off between the proximity to the true biologic process and the concept of parsimony.

Once we have defined the trajectory function, we can connect it with the model for the observed values $\mathbf{Y}_i = (Y_i(t_{i1}), Y_i(t_{i2}), \ldots, Y_i(t_{iJ_i}))$. It can be done as follows:

$$\begin{aligned} Y_i(t_{ij}) &= W_i(t_{ij}) + \epsilon_i(t_{ij}),\ i = 1, 2, \ldots, n \text{ and } j = 1, 2, \ldots, J_i \\ &= \mathbf{x}_i\boldsymbol{\beta} + \mathbf{f}(t_{ij})^\top \mathbf{b}_i + \epsilon_i(t_{ij}), \end{aligned} \tag{2.8}$$

where $Y_i(t_{ij})$ is a random variable representing the observed value of the longitudinal variable at time t_{ij}, and $W(t_{ij})$ is the trajectory function evaluated at this same time. It represents the true - and not necessarily observed - value of the longitudinal variable. Here, the trajectory function is composed of both fixed and random effects. Then, $\mathbf{x}_i$ is the vector of covariates for the i-th subject, and $\boldsymbol{\beta}$ is the vector of coefficients associated with the covariates. The variation with time of this variable is represented by the vector $\mathbf{f}(.)$ and $\mathbf{b}_i$ is the vector of random effects. We will consider this vector as normally distributed with mean $\boldsymbol{\mu}_b$ and variance-covariance matrix Σ_b. This is a basic and usual assumption in this framework. At last, $\epsilon_i(t_{ij}) \sim Normal(0, \sigma_\epsilon^2)$ is the measurement error for the i-th subject at the j-th time point. We call attention to the fact that the marginal distribution of $Y_i(t_{ij})$ is also Normal, and its parameters are described on the Appendix A, page 99. In addition to that, by using the formulation in Equation (2.8) to connect the trajectory function and the observed values, we are assuming that exists a significant and non-ignorable measurement error. If such assumption does not match with the goals of the study, we can simply ignore the error term ϵ_i.

In order to make these concepts clearer, we prepared a toy example. Then, Figure 1 shows an illustration of the trajectory function on a framework where the measurements were prone to error. The black "x" marks are equivalent to the observed values of the longitudinal and time-dependent covariate, which is assumed to be measured with error. So, on this example, there were six measurements along the follow-up period of this hypothetical study. The time points were $\mathbf{t}_i = (0, 2, \ldots, 10)$. In addition, the continuous red line is a reference to the trajectory function, representing the true and, in this case, unobserved values of the longitudinal covariate.

Turning our attention to the survival component, we can use an adaptation of the proportional hazards model for time-dependent covariates (Tsiatis and Davidian, 2004). This model was described in Section 2.2.1 (page 26) and the extension is given by:

$$\begin{aligned} h(u_i) &= \lim_{du \to 0} \frac{\mathbb{P}(u_i \leq T_i < u_i + du | T_i \geq u_i, W_i^H(u_i), \mathbf{z}_i)}{du} \\ &= h_0(u_i) \exp\left\{\eta W_i(u_i) + \mathbf{z}_i \boldsymbol{\psi}\right\}, \; i = 1, 2, \ldots, n, \end{aligned} \tag{2.9}$$

where u_i is the observed survival time of subject i. The random variable T_i represents the time to failure. Then, $W_i^H(u_i) = \{W_i(t),\, 0 \leq t < u_i\}$ is the entire history of the longitudinal variable up to time u_i. The covariates vector for this sub-model in represented by $\mathbf{z}_i$, and $\boldsymbol{\psi}$ is the corresponding vector of coefficients. In turn, $h_0(t)$ is the baseline hazard function evaluated on time u_i. The parameter that links both sub-models is η; this is a very important quantity in this framework. It informs if there is an actual relationship between the longitudinal and the survival variables. In case there is, it also tells the strength of

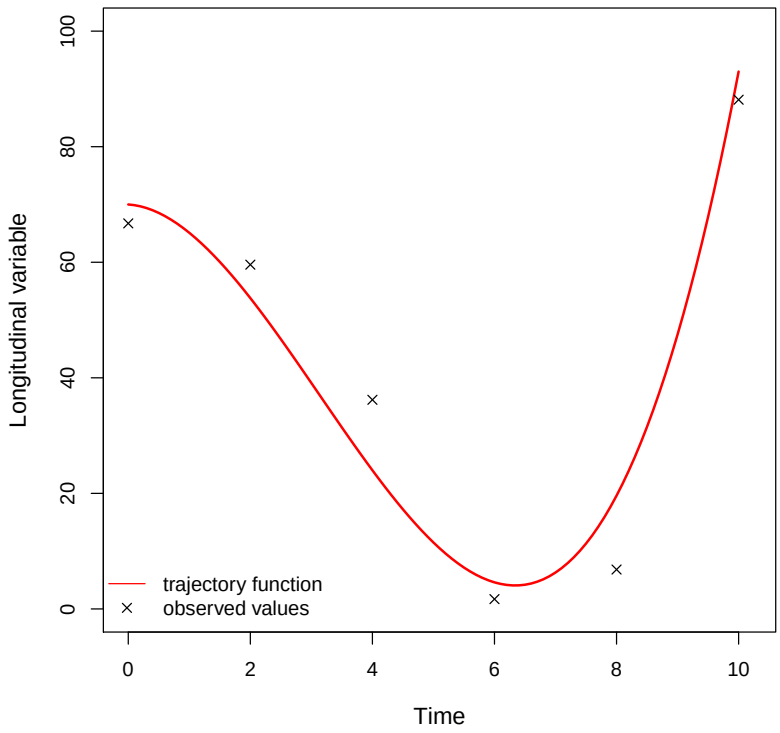


Figure 1 – Illustration of the difference between the trajectory function and the observed values.

this relationship.

We highlight that we use the "true" estimated value $W_i(u_i)$ in the proportional hazards model for time-dependent covariates in Equation (2.9). Then, the vector of random effects, which is a part of the trajectory function, is also in the survival sub-model. The reason for using $W_i(u_i)$ is because there is no guarantee that we will observe the longitudinal variable at this time point, that is, $y_i(u_i)$. In the case we do observe it, this value can be measured with error. Moreover, it is also noteworthy that the model formulation for the hazard function in (2.9) is conditioned on the entire history of the longitudinal variable $W_i^H(u)$. However, at the end, out of the entire history, we use only the value at the observed survival time u_i. Nonetheless, we do use this information $W_i^H(.)$ on the survival function (Rizopoulos, 2012), as it is shown on Equation (2.10).

$$\begin{aligned} S(u_i) &= \exp\{-H(u_i)\} = \exp\left\{-\int_0^{u_i} h(s)ds\right\} \\ &= \exp\left\{-\int_0^{u_i} h_0(s)\exp\left\{\eta W_i(s) + \mathbf{z}_i\boldsymbol{\psi}\right\}ds\right\} \end{aligned} \tag{2.10}$$

Finally, we have completely specified both longitudinal and survival components. These specifications are composed by the Equations (2.7), (2.8) and (2.9). Then, we can

write the joint likelihood function:

$$
\begin{aligned}
L(\boldsymbol{\Phi}; Data) &= \prod_{i=1}^{n} p(u_i, \mathbf{y}_i|\boldsymbol{\Phi}) = \prod_{i=1}^{n} \int p(u_i, \mathbf{y}_i, \mathbf{b}_i|\boldsymbol{\Phi}) d\mathbf{b}_i = \prod_{i=1}^{n} \int p(u_i, \mathbf{y}_i|\mathbf{b}_i, \boldsymbol{\Phi}) p(\mathbf{b}_i|\boldsymbol{\mu}_b, \Sigma_b) d\mathbf{b}_i \\
&= \prod_{i=1}^{n} \int \prod_{j=1}^{J_i} p(u_i, y_i(t_{ij})|\mathbf{b}_i, \boldsymbol{\Phi}) p(\mathbf{b}_i|\boldsymbol{\mu}_b, \Sigma_b) d\mathbf{b}_i \\
&= \prod_{i=1}^{n} \int \prod_{j=1}^{J_i} p(u_i|y_i(t_{ij}), \mathbf{b}_i, \boldsymbol{\Phi}) p(y_i(t_{ij})|\mathbf{b}_i, \boldsymbol{\Phi}) p(\mathbf{b}_i|\boldsymbol{\mu}_b, \Sigma_b) d\mathbf{b}_i \\
&= \prod_{i=1}^{n} \int h(u_i|\boldsymbol{\Phi})^{\delta_i} \exp\left(-H(u_i|\boldsymbol{\Phi})\right) \left(\prod_{j=1}^{J_i} p(y_i(t_{ij})|\mathbf{b}_i, \boldsymbol{\Phi}) \right) p(\mathbf{b}_i|\boldsymbol{\mu}_b, \Sigma_b) d\mathbf{b}_i \\
&= \prod_{i=1}^{n} \int \left[h_0(u_i) \exp\left(\eta W_i(u_i) + \mathbf{z}_i\boldsymbol{\psi}\right)\right]^{\delta_i} \exp\left\{ - \int_0^{u_i} h_0(s) \exp\left(\eta W_i(s) + \mathbf{z}_i\boldsymbol{\psi}\right) ds \right\} \\
&\quad \frac{1}{(2\pi\sigma_\epsilon^2)^{J_i/2}} \exp\left\{ - \sum_{j=1}^{J_i} \frac{\{Y_i(t_{ij}) - W_i(t_{ij})\}^2}{2\sigma_\epsilon^2} \right\} p(\mathbf{b}_i|\boldsymbol{\mu}_b, \Sigma_b) d\mathbf{b}_i, \qquad (2.11)
\end{aligned}
$$

where $\boldsymbol{\Phi} = (\boldsymbol{\beta}, \boldsymbol{\mu}_b, \Sigma_b, \eta, \boldsymbol{\psi}, \sigma_\epsilon^2)$ is the vector of all unknown quantities to be estimated. $Data$ represents the observed data. In this case we have that $Data = \{u_i, \delta_i, \mathbf{y}_i, \mathbf{t}_i, \mathbf{x}_i, \mathbf{z}_i,$ for $i = 1, 2, \ldots, n\}$. Then for the i-th subject, u_i is the observed survival time and δ_i is the indicator of censorship/failure. The vector of the observed measurements for the i-th individual is $\mathbf{y}_i = (y_i(t_{i1}), y_i(t_{i2}), \ldots, y_i(t_{iJ_i}))$; $\mathbf{t}_i = (t_{i1}, t_{i2}, \ldots, t_{iJ_i})$ is the vector of time points in which the measurements were collected, and J_i is the number of measurements. The vector of covariates that are possibly associated with the longitudinal variable is represented by $\mathbf{x}_i$ and $\boldsymbol{\beta}$ is the vector of coefficients associated with these covariates. In the survival sub-model the vector of covariates for the i-th subject is $\mathbf{z}_i$ and the vector of coefficients is $\boldsymbol{\psi}$. It is worth mentioning that the vectors of covariates $\mathbf{x}_i$ and $\mathbf{z}_i$ are not necessarily the same. The measurement error is normally distributed with mean 0 and variance σ_ϵ^2. The parameter η is the one that links both sub-models. Next, the vector of random effects at individual level is $\mathbf{b}_i$, for $i = 1, 2, \ldots, n$, with n representing the sample size. As we assume that this vector is normally distributed, $\boldsymbol{\mu}_b$ represents its mean and Σ_b is the variance-covariance matrix. The vector of true values for the longitudinal variable is represented by $\mathbf{W}_i = (W_i(t_{i1}), W_i(t_{i2}), \ldots, W_i(t_{iJ_i}))$, for the i-th subject. This trajectory function is described in details in Equations (2.7) and (2.8). At last, the baseline hazard function is $h_0(.)$, the hazard function is $h(.)$, and its cumulative function is $H(.)$.

Besides the complexity of joint models, a potential difficulty when dealing with this approach is that we have to solve a complicated integral in Equation (2.11). The complication of this calculation is due to the fact that there are two functions that depend on the time. These functions are the baseline hazard $h_0(.)$ and the time-varying part of the trajectory function $W_i(.)$. So, in a small set of combinations of these two functions, it is possible to solve this integral analytically. Usually, these combinations are composed by

simple structures. Then, in a sense, this would be a restriction.

A solution for this challenge is to solve this integral via Gaussian quadrature (Rizopoulos, 2012). In this case, we can choose complex forms for both time-dependent functions. In a very brief explanation, this method approximates the integral by transforming it into summations. In this sum, we need a vector of Q nodes and a vector of weights with the same size. Here, Q represents the number of quadratures. In turn, the vector of nodes represents in which points we will calculate the function that we want to integrate. Then, the resulting approximation is given by the calculation we just mentioned weighted by the vector of weights. There are several forms of quadratures and one can find more information about them in Kahaner et al. (1989). We point out that, since this method is an approximation, it is evident that it presents estimates with error. However, there are several quadrature methods and they usually present good approximations.

The method we will use to work this obstacle around is the Gauss-Kronrod quadrature with $Q = 15$ nodes. This specific method was also used by Brilleman et al. (2017). The nodes and the weights are originally calculated for the interval $(-1, 1)$. Nonetheless, we can apply a linear transformation to obtain the vector nodes at the interval we are interested in, which is $(0, u_i)$, for each subject $i = 1, 2, \ldots, n$.

An additional comment is that this specific form of quadrature for $Q = 7$ and $Q = 15$ is available in the `R` (R Core Team, 2019) package `pracma` (Borchers, 2019). The command is `gauss_kronrod`. The vector of nodes and weights in the original scale for each specification of quadratures is also available within this function. Even so, we implemented it ourselves.

At last the Gauss-Kronrod approximation for the integral in Equation (2.11) is given by

$$\begin{aligned} H(u_i) &= \int_0^{u_i} h(t)dt \\ &\approx \frac{u_i}{2}\sum_{q=1}^{Q} w_q h\left(\frac{u_i(1+t_q)}{2}\right) \\ &\approx \frac{u_i}{2}\sum_{q=1}^{Q} w_q h_0\left(\frac{u_i(1+t_q)}{2}\right)\exp\left\{\mathbf{z}_i\boldsymbol{\psi} + \eta W_i\left(\frac{u_i(1+t_q)}{2}\right)\right\} \\ &\approx \frac{u_i}{2}\exp\left\{\mathbf{z}_i\boldsymbol{\psi}\right\}\sum_{q=1}^{Q} w_q h_0\left(\frac{u_i(1+t_q)}{2}\right)\exp\left\{\eta W_i\left(\frac{u_i(1+t_q)}{2}\right)\right\}, \end{aligned} \tag{2.12}$$

where w_q represents the q-th component of the weights vector, and t_q is the q-th reescaled time point in which we will calculate the functions $h_0(.)$ and $W_i(.)$.

In the next chapter, we will discuss the Bernstein Polynomials. This is the method we use to approximate both baseline hazard function and the time-varying aspect of the

longitudinal variable under the joint model framework. The usage of the BP in this manner is one of the main contributions of the present thesis.

3 Bernstein Polynomials

The main goal of this chapter is to present and discuss the so called Bernstein Polynomials, which is one of the key topics and the central contribution of this thesis. The general usage of the BP is to approximate any smooth curve/function. An advantage of approximating functions through polynomials is that they can be specified in a simple way since they can be written in the form of summations; thereby, calculations of derivatives and gradient matrices are relatively easy to obtain (Osman and Ghosh, 2012). In addition, polynomials are infinitely differentiable (de Figueiredo, 1996). This feature allows us to analyze the approximation with details. In this chapter, we will also explore some of the mathematical and applied aspects of this methodology. A crucial discussion involving polynomials is about their order. The degree/order of the BP plays an important role in the approximation performance. Therefore, we aim to provide a way of fixing a minimum degree so that important aspects of the target function can remain in the approximation. We also propose two robust methods to serve as a guidance in the specification of the maximum degree.

Bernstein Polynomials were proposed by Sergei Natanovich Bernstein in 1912. The idea that led to the development of BP arose from a demonstration of a special case of the Weierstrass theorem (Lorentz, 1986; Bernstein, 1912). This theorem is formally enunciated in the following way (Bartle and Sherbert, 2011):

Theorem 1 (Weierstrass Approximation Theorem)**.** *Let $I = [a, b]$ and let $f : I \to \mathbb{R}$ be a continuous function. If $\varepsilon > 0$ is given, then there exists a polynomial function p_ε such that $|f(x) - p_\varepsilon(x)| < \varepsilon$ for all $x \in I$.*

The implication of this theorem is that it is possible to approximate any continuous function f, defined on the closed interval $[a, b]$ by a polynomial p_ε. In other words, for every value x in the interval $I = [a, b]$, the absolute difference between $f(x)$ and the polynomial p_ε is lower than a pre-specified value $\varepsilon > 0$, and ε can be as small as desired. As it was stated by Bartle and Sherbert (2011), mathematically, a high degree of the polynomial should be considered in order to obtain a good approximation for the target function f. This function f can be a density function, cumulative distribution function, or a mean curve, for example.

The demonstration given by Bernstein was constructed based on the theory of probabilities and his idea was the following: consider an event A such that the probability that it happens is equal to x, *i. e.* $\mathbb{P}(A) = x$. Next, assume that m trials of this experiment are performed in a way that the quantity $f\left(k/m\right)$ will be payed to a hypothetical player if the event A occurs k times. Then, define a random variable K as the number of successes

(occurrence of the event A) in m trials. It is clear that this random variable follows the $Binomial(m, x)$ distribution. Therefore, the probability that the event A will occur k times in m trials is equal to

$$\mathbb{P}(K = k) = \binom{m}{k} x^k (1-x)^{m-k},\ k = 0, 1, \ldots, m$$

and the expected value of the quantity that is going to be paid to the player in this situation is

$$E_m(x) = \sum_{k=0}^{m} f\left(\frac{k}{m}\right) \binom{m}{k} x^k (1-x)^{m-k}. \tag{3.1}$$

Based on what was exposed above, Bernstein demonstrated that $\sup_{x\in[0,1]} (|f(x) - E_m(x)|) < \varepsilon$, for an $\varepsilon > 0$ (Bernstein, 1912). In other words, $E_m(x)$ converges to $f(x)$ as $m \to \infty$, that is:

$$f(x) = \lim_{m\to\infty} \sum_{k=0}^{m} f\left(\frac{k}{m}\right) \binom{m}{k} x^k (1-x)^{m-k}.$$

The BP approximation of order m for the function f is given by $E_m(x)$ in Equation (3.1). The part $b_{k,m}(x) = \binom{m}{k} x^k (1-x)^{m-k}$, $k = 0, 1, \ldots, m$ is called the Bernstein basis and m is the degree of this polynomial. We will often represent the vector of Bernstein basis by $\mathbf{b}_m(x) = (b_{0,m}(x), b_{1,m}(x), \ldots, b_{m,m}(x))^\top$. We highlight that a BP with degree m is composed of $m + 1$ components. More details related to the demonstration of the convergence can be found in Bernstein (1912), Kuller (1964) and Lorentz (1986).

It is noteworthy that Equation (3.1) can be rewritten in the form of a Beta density. In this thesis we will use both ways of writing the Bernstein basis:

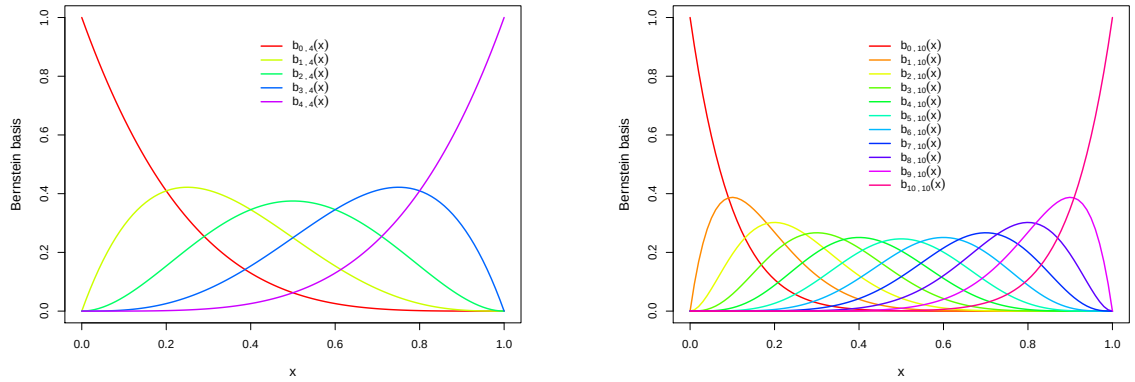
$$b_{k,m}(x) = \binom{m}{k} x^k (1-x)^{m-k} = \frac{1}{m+1} f_{Beta}(x; k+1, m-k+1),$$

where $f_{Beta}(x; a, b)$ represents the density function of a $Beta(a, b)$ distribution evaluated at the point $x \in (0, 1)$.

In addition, the Bernstein basis can be seen as weights in (3.1), since $0 \leq b_{k,m}(x) \leq 1$ for all $k = 0, 1, \ldots, m$, as it represents probabilities of the Binomial distribution for all possible number of successes. For the same reason, the vector of Bernstein basis also sums up to 1:

$$\sum_{k=0}^{m} b_{k,m}(x) = \sum_{k=0}^{m} \binom{m}{k} x^k (1-x)^{m-k} = 1.$$

An illustration of the vector of Bernstein basis for $m = 4$ and for $m = 10$ can be seen in Figure 2a and Figure 2b, respectively. Note that we have $m + 1$ lines in each figure.



(a) $m = 4$. (b) $m = 10$.

Figure 2 – Illustration of the vector of Bernstein basis for $m = 4$ and $m = 10$.

As mentioned above, the vector of basis has a role as weights. It can be noticed in both panels (2a and 2b) that the weights varies with x, which is the probability of the Binomial distribution. Thus, the approximation of the target function $f(.)$ will be weighted by $m + 1$ values coming from the vector of basis. Clearly, when $m = 10$ we will have more information to weight the values of the function, resulting in a more accurate approximation.

Bernstein (1912) highlights that the approximation via BP of the function f requires the knowledge of values (or approximated values) of the very function f. That is, analytically, in order to approximate the desired function by the BP, it is not necessary to know the behavior of this function, nor most of its characteristics. It is only required that f is continuous and to know the values of f on $m + 1$ specific points of the domain, where m represents the degree of the BP. These points usually are the set $\{k/m : k = 0, 1, \ldots, m\}$.

Other interesting feature of the BP is the fact it maintains the shape restriction of the target functions with simple constraints on its formulation (Chang et al., 2007; Osman and Ghosh, 2012; Wang and Ghosh, 2012, 2013). These restrictions may be, for example, monotone, concave/convex, increasing/decreasing behaviors. This is very convenient, since the hazard function assumes values greater than 0, and growth curves and cumulative hazard function are non-decreasing functions, as examples.

Based on what has been discussed above, the next two sections will concentrate on explaining how to use the BP to model functions of the longitudinal and survival

components of the joint model.

3.1 Bernstein Polynomials to Model the Longitudinal Component

In the previous section it was explained that it is possible to use the BP to obtain an approximation for any function $f(.)$, as long as this function is continuous and it is defined in the interval $(0, 1)$. Furthermore, it was discussed in Section 2.1.1 that a reasonable option to model the longitudinal variable is through Linear Mixed Effects Model (LME). The structure of the LME model is composed of both fixed and - possibly normally distributed - random effects. We will use the Mixed Effect structure to approximate the time-varying aspect of the longitudinal variable for an interval $(0, T_{max})$ via Bernstein Polynomials with degree $m_L - 1$. This approach was based on the proposal of Wang and Ghosh (2013). For clarification, we define T_{max} as the lower integer greater than the maximum of all observed measurement and survival times, *i. e.*, $T_{max} = \left\lceil \max_{\substack{i\in\{1,2,\ldots,n\}\\ j\in\{1,2,\ldots J_i\}}} (t_{ij}, u_i) \right\rceil$.

Wang and Ghosh (2013) proposed a procedure to model the variation with respect to the time of the longitudinal variable using the Bernstein Polynomials with degree $m_L - 1$. In their proposal for longitudinal data alone, the information from the data consisted only of the observed measurements (possibly with error) and the time point in which each measurement was taken. Thus, the role of the BP in their approach was to provide an approximation for the trajectory function $W(.)$. The mentioned structure was the following

$$\begin{aligned} Y_i(t_{ij}) &= W_i(t_{ij}) + \epsilon_i(t_{ij}),\ i = 1, 2, \ldots n,\ j = 1, 2, \ldots, J_i \\ &= \sum_{l=1}^{m_L} f_i\left(\frac{l-1}{m_L-1} T_{max}\right) \binom{m_L-1}{l-1} \left(\frac{t_{ij}}{T_{max}}\right)^{l-1} \left(1 - \frac{t_{ij}}{T_{max}}\right)^{m_L-l} + \epsilon_i(t_{ij}) \\ &= \sum_{l=1}^{m_L} \xi_{i,l}^{m_L-1} b_{l,m_L-1}\left(\frac{t_{ij}}{T_{max}}\right) + \epsilon_i(t_{ij}) \\ &= (\boldsymbol{\xi}_i^{m_L-1})^\top \mathbf{b}_{m_L-1}\left(\frac{t_{ij}}{T_{max}}\right) + \epsilon_i(t_{ij}), \end{aligned} \tag{3.2}$$

where $Y_i(t_{ij})$ is a random variable referring to the observed value of the longitudinal variable for the i-th subject and the j-th measurement time, $W_i(t_{ij})$ represents the true value of this variable, and $\epsilon_i(t_{ij})$ is the measurement error. Here, $\epsilon_i(t_{ij}) \sim Normal(0, \sigma_\epsilon^2)$. Suppose that the temporal evolution of the longitudinal variable is in accordance with a function f_i and that it will be modeled by Bernstein Polynomials with degree m_L - 1. Thus, $\boldsymbol{\xi}_i^{m_L-1} = (\xi_{i,1}^{m_L-1}, \xi_{i,2}^{m_L-1}, \ldots, \xi_{i,m_L}^{m_L-1})^\top$ is the vector composed of m_L Bernstein coefficients, at subject-level, approximating the values of the function $f_i(.)$ for each subject i at time points $[(l-1)/(m_L-1)]T_{max}$, $l = 1, 2, \ldots, m_L$. At last, $\mathbf{b}_{m_L-1}\left(\frac{t_{ij}}{T_{max}}\right) =$

$\left(b_{1,m_L-1}\left(\frac{t_{ij}}{T_{max}}\right), b_{2,m_L-1}\left(\frac{t_{ij}}{T_{max}}\right), \ldots, b_{m_L,m_L-1}\left(\frac{t_{ij}}{T_{max}}\right)\right)^\top$ is the vector composed of m_L Bernstein basis that will weight the information from the vector of coefficients.

As we will include both fixed and random effects - which was not the case of Wang and Ghosh (2013) -, the proposed longitudinal sub-model using the BP will be

$$\begin{aligned} Y_i(t_{ij}) &= W_i(t_{ij}) + \epsilon_i(t_{ij}),\ i = 1, 2, \ldots n\ ,\ j = 1, 2, \ldots, J_i \\ &= \mathbf{x}_i\boldsymbol{\beta} + (\boldsymbol{\xi}_i^{m_L-1})^\top \mathbf{b}_{m_L-1}\left(\frac{t_{ij}}{T_{max}}\right) + \epsilon_i(t_{ij}), \end{aligned}$$

where $\mathbf{x}_i$ is the vector of covariates for the i-th subject that are possibly related to the longitudinal variable, and $\boldsymbol{\beta}$ is the vector of coefficients associated with these covariates, for $i = 1, 2, \ldots, n$.

Wang and Ghosh (2013) assumed that the vector of coefficients of the BP was normally distributed, that is, $\boldsymbol{\xi}_i^{m_L-1} \sim N_{m_L}(\boldsymbol{\mu}_{\boldsymbol{\xi}}, \Sigma_{\boldsymbol{\xi}})$, for $i = 1, 2, \ldots, n$. In this case, $\boldsymbol{\mu}_{\boldsymbol{\xi}} = (\mu_{\xi_1}, \mu_{\xi_2}, \ldots, \mu_{\xi_{m_L}})^\top$ represents the overall mean of the longitudinal variable changing with time and $\Sigma_{\boldsymbol{\xi}}$ is the $(m_L \times m_L)$ variance-covariance matrix, accommodating the correlation coming from the different measurements of the same subject. As a result, $W_i(.)$, which represents the true value of the longitudinal variable, is approximated by a Gaussian Process. We call attention to the fact that there is a straight relationship between the vector of coefficients $\boldsymbol{\xi}_i^{m_L-1}$ and the function being approximated; see Equation (3.2). Thus, if necessary it is possible to impose constraints to this vector to maintain essential characteristics of the target function.

Since the measurement error $\epsilon_i(t_{ij})$ follows a Normal distribution, it is clear that the distribution of $Y_i(t_{ij})$ conditioned on the vector of random effects (whether they are coefficients of the Bernstein Polynomials or the regular random intercept and slope) also follows a Normal distribution. That is, $Y_i(t_{ij})|\boldsymbol{\xi}_i^{m_L-1} \sim Normal\left(\mathbf{x}_i\boldsymbol{\beta} + (\boldsymbol{\xi}_i^{m_L-1})^\top \mathbf{b}_{m_L-1}\left(\frac{t_{ij}}{T_{max}}\right), \sigma_\epsilon^2\right)$, for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, J_i$. Moreover, the marginal distribution of $Y_i(t_{ij})$ is Normal with mean and variance given, respectively, by

$$\mathbb{E}[Y_i(t_{ij})] = \mathbb{E}\left[\mathbb{E}[Y_i(t_{ij})|\boldsymbol{\xi}_i^{m_L-1}]\right] = \mathbf{x}_i\boldsymbol{\beta} + (\boldsymbol{\mu}_\xi^{m_L-1})^\top \mathbf{b}_{m_L-1}\left(\frac{t_{ij}}{T_{max}}\right), \tag{3.3}$$

and

$$\begin{aligned} Var[Y_i(t_{ij})] &= \mathbb{E}\left[Var\left[Y_i(t_{ij})|\boldsymbol{\xi}_i^{m_L-1}\right]\right] + Var\left[\mathbb{E}\left[Y_i(t_{ij})|\boldsymbol{\xi}_i^{m_L-1}\right]\right] \\ &= \sigma_\epsilon^2 + \sum_{k=1}^{m_L}\sum_{l=1}^{m_L} b_{k,m_L-1}\left(\frac{t_{ij}}{T_{max}}\right) b_{l,m_L-1}\left(\frac{t_{ij}}{T_{max}}\right)\sigma_{kl}, \end{aligned} \tag{3.4}$$

where σ_{kl} is the covariance between the coefficients $\xi_{i,k}^{m_L-1}$ and $\xi_{i,l}^{m_L-1}$, for all i. Note that both overall mean and variance may change along the time. These results can be seen with more details in the Appendix A (page 99).

An important restriction is that the degree $m_L - 1$ of the BP in this specific approach must be greater than 1. Besides that, $m_L - 1$ has to be lower than the maximum number of measurements. According to Wang and Ghosh (2013), the first requirement is made aiming to obtain a non-degenerated Gaussian Process, while the second one is to avoid collinearity issues. Thus, the degree of the BP for the longitudinal sub-model is an integer in the interval $m_L \in [2, \max_i J_i)$, where J_i is the number of measurements for each one of the n subjects under study.

Considering the longitudinal sub-model, the unknown quantities to be estimated are $\boldsymbol{\beta}$, $\boldsymbol{\mu_\xi}$, $\Sigma_{\boldsymbol{\xi}}$ and σ_ϵ^2. Since we are basing the inference process under a Bayesian point of view, prior distributions are specified for these parameters. Moreover, it is highlighted here that, in case it is desired and necessary, it is also possible to shape constraint the BP by imposing restrictions on the vector of overall means $\boldsymbol{\mu_\xi}$.

It is worth pointing out that $\xi_{i,l}^{m_L-1} \approx f_i([(l-1)/(m_L-1)]T_{max})$ for $l = 1, 2, \ldots, m_L$. Then, we will thin the grid of points at which the function $f_i(.)$ will be evaluated - within the interval $(0, T_{max})$ - by making $m_L \to \infty$. Thus, in a way, the greater is the degree, the more accurate will be the estimation of the function $f_i(.)$. In addition to that, there may have a case where we will have estimates for this function in contradictory time points, *i. e.*, time points after which a specific subject is no longer being followed-up. This happens because the BP smooths the target function by estimating it in all the domain $(0, T_{max})$, but putting lower weights as time gets far from the observed measurement times t_{ij}, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, J_i$. Therefore, this method does give estimates for times that may not make sense in practice, since there will be estimates for all time points $[(l-1)/(m_L-1)]T_{max}$ even for those subjects who had stopped being followed-up before one of these points, because of death, for example. The idea is to take into account values that were observed as well as those that could have been observed, with the latter having smaller probability (or weight). A better discussion for this matter is given in Section 3.3. The next section focuses on explaining how to approximate both cumulative baseline hazard function and baseline hazard function via Bernstein Polynomials.

3.2 Bernstein Polynomials to Model the Survival Component

This section focuses on the details of the survival sub-model. This sub-model was defined as the proportional hazard model for time dependent covariates. Our proposal is to approximate both cumulative baseline hazard function $H_0(.)$ and baseline hazard function $h_0(.)$ using Bernstein Polynomials. The scheme is to first approximate $H_0(.)$ and

then, the approximation of $h_0(.)$ is obtained straightforwardly. Our main interest relies on the baseline hazard function, since it is directly a part of the proportional hazards model (with or without time-dependent variables). This modeling approach was based on the ideas seen in Osman and Ghosh (2012), with the difference that their main goal relied on the case where survival curves cross each other along the time. So, our aim in this section is to explain how to use this procedure on the survival sub-model in the joint model framework.

First, consider the approximation of the Bernstein Polynomials with degree m_S for the cumulative baseline hazard function H_0. As defined in Equation (3.1) this approximation is given by:

$$\begin{aligned} H_0(u; m_S) &\approx \sum_{k=0}^{m_S} H_0\left(\frac{k}{m_S} T_{max}\right) \binom{m_S}{k} \left(\frac{u}{T_{max}}\right)^k \left(1 - \frac{u}{T_{max}}\right)^{m_S - k} \\ &= \sum_{k=1}^{m_S} H_0\left(\frac{k}{m_S} T_{max}\right) \binom{m_S}{k} \left(\frac{u}{T_{max}}\right)^k \left(1 - \frac{u}{T_{max}}\right)^{m_S - k} \\ &= \sum_{k=1}^{m_S} H_0\left(\frac{k}{m_S} T_{max}\right) b_{k,m_S}\left(\frac{u}{T_{max}}\right), \end{aligned} \tag{3.5}$$

where m_S is the degree of the BP for the survival sub-model, T_{max} is fixed as a value greater than the maximum of the observed survival times and $b_{k,m_S}(u/T_{max})$ is k-th the Bernstein basis, for $k = 0, 1, \ldots, m_S$. The value T_{max} should be in accordance with this quantity in the approximation of the BP in the longitudinal sub-model, since the time scale is the same. It is guaranteed by the Weierstrass theorem that there is uniform convergence on the interval $(0, T_{max}]$ as $m_S \longrightarrow \infty$ (Bernstein, 1912; Lorentz, 1986; Osman and Ghosh, 2012). That is, $H_0(u) = \lim_{m_S \to \infty} H_0(u; m_S)$.

The approximation for the baseline hazard function $h_0(.)$ via BP is obtained by taking the derivative of expression (3.5) above. Thus,

$$\begin{aligned} h_0(u; m_S) &= \frac{\partial}{\partial u} H_0(u; m_S) \\ &= \sum_{k=1}^{m_S} \left[H_0\left(\frac{k}{m_S} T_{max}\right) - H_0\left(\frac{k-1}{m_S} T_{max}\right) \right] \frac{m_S}{T_{max}} \binom{m_S - 1}{k - 1} \left(\frac{u}{T_{max}}\right)^{k-1} \left(1 - \frac{u}{T_{max}}\right)^{m_S - k} \\ &= \sum_{k=1}^{m_S} \left[H_0\left(\frac{k}{m_S} T_{max}\right) - H_0\left(\frac{k-1}{m_S} T_{max}\right) \right] \frac{m_S}{T_{max}} f_{Binomial}\left(k - 1; m_S - 1, \frac{u}{T_{max}}\right) \\ &= \sum_{k=1}^{m_S} \left[H_0\left(\frac{k}{m_S} T_{max}\right) - H_0\left(\frac{k-1}{m_S} T_{max}\right) \right] \frac{f_{Beta}\left(u/T_{max}; k, m_S - k + 1\right)}{T_{max}}, \end{aligned} \tag{3.6}$$

where $f_{Binomial}\left(k - 1; m_S - 1, u/T_{max}\right)$ represents the probability of $k - 1$ successes from a $Binomial(m_S - 1, u/T_{max})$ distribution and $f_{Beta}\left(u/T_{max}; k, m_S - k + 1\right)$ is a density function of a $Beta(k, m_S - k + 1)$ distribution evaluated at the point u/T_{max}. It is also

true, by the Weierstrass theorem, that $h_0(.; m_S) \longrightarrow h_0(.)$ uniformly on the interval $(0, T_{max}]$ as $m_S \longrightarrow \infty$. We point out that the approximation for the baseline hazard function is via a BP with degree $m_S - 1$, since the basis function is a probability of a $Binomial(m_S - 1, u/T_{max})$. More details about this result can be found in the Appendix A (page 100).

Note from Equations (3.5) and (3.6) that it is necessary to know the values of the function H_0 at $m_S + 1$ points of the domain $(0, T_{max})$. However, since the function H_0 is unknown, this is not a realistic requirement in the present scenario. Thus, in order to work this obstacle around, we will simply represent the differences between the cumulative baseline hazard functions as a vector of parameters to be estimated. That is,

$$\begin{aligned} h_0(u; m_S) &\approx \sum_{k=1}^{m_S} \left[H_0\left(\frac{k}{m_S} T_{max}\right) - H_0\left(\frac{k-1}{m_S} T_{max}\right) \right] \frac{f_{Beta}\left(u/T_{max}; k, m_S - k + 1\right)}{T_{max}} \\ &= \sum_{k=1}^{m_S} \gamma_k^{m_S - 1} \frac{f_{Beta}\left(u/T_{max}; k, m_S - k + 1\right)}{T_{max}}. \end{aligned} \tag{3.7}$$

Therefore, $\gamma_k^{m_S-1}$, for $k = 1, 2, \ldots, m_S$, will represent the difference between the unknown cumulative baseline hazard function applied at two different points. It is important to mention that $\gamma_k^{m_S-1} \approx H_0\left(\frac{k}{m_S} T_{max}\right) - H_0\left(\frac{k-1}{m_S} T_{max}\right) \geq 0$ since $H_0\left(\frac{k}{m_S} T_{max}\right) \geq H_0\left(\frac{k-1}{m_S} T_{max}\right)$. Then, it is assured that the model for the cumulative baseline hazard function in (3.5) provides a non-decreasing approximation. In other words, the BP approximation for the $H_0(.)$, based on the work of Osman and Ghosh (2012), is in agreement with one important property of this function.

An interesting fact is that the difference between the points in which the function H_0 is evaluated (Equation (3.7)) diminishes as the degree of the BP increases, *i. e.*, $\frac{k}{m_S} T_{max} - \frac{(k-1)}{m_S} T_{max} = \frac{1}{m_S} T_{max} \xrightarrow[m_S \to \infty]{} 0$. Therefore,

$$\gamma_k^{m_S-1} = H_0\left(\frac{k}{m_S} T_{max}\right) - H_0\left(\frac{k-1}{m_S} T_{max}\right) = \int_{(k-1)T_{max}/m_S}^{(k/m_S)T_{max}} h_0(u)du \xrightarrow[m_S \to \infty]{} h_0\left(\frac{k}{m_S} T_{max}\right),$$

since there is no accumulation of the hazards when $m_S \to \infty$. So, the greater the degree m_S is, the more accurate will be the estimates of the vector $\boldsymbol{\gamma}_{m_S-1}$ in the sense of having values of the target function $h_0(.)$ over a fine grid of the domain . This is also true for the cumulative hazard function. Moreover,

$$H_0\left(\frac{k}{m_S} T_{max}\right) \approx \sum_{o=1}^{k} \gamma_o^{m_S-1}, \tag{3.8}$$

for $k = 1, 2, \ldots, m_S$, precisely because, summing up the components of $\boldsymbol{\gamma}_{m_S-1}$ is equivalent to accumulating values of the baseline hazard function (see more about this result in the Appendix A, page 102).

The proposed model for the survival component is fully specified. For clarification, we remind that we will use the BP with degree m_S to approximate the cumulative baseline hazard function under the proportional hazards model for time dependent covariates (Equation (2.9)). Thus, we will use the whole procedure described in this section as

$$\begin{aligned} h(u_i) &= h_0(u_i) \exp\left\{\eta W_i(u_i) + \mathbf{z}_i^\top \boldsymbol{\psi}\right\}, \; i = 1, 2, \ldots, n, \\ &= \left[\sum_{k=1}^{m_S} \gamma_k^{m_S-1} \frac{f_{Beta}\left(u/T_{max}; k, m_S - k + 1\right)}{T_{max}}\right] \exp\left\{\eta W_i(u_i) + \mathbf{z}_i^\top \boldsymbol{\psi}\right\}. \end{aligned}$$

In addition, as we described on page 35, the cumulative hazard function will be approximated by the Gauss-Kronrod quadrature with $Q = 15$ nodes. Then, in this case, we have

$$\begin{aligned} H(u_i) &= \int_0^{u_i} h(s)ds = \int_0^{u_i} h_0(s) \exp\left\{\eta W_i(s) + \mathbf{z}_i^\top \boldsymbol{\psi}\right\} ds \\ &\approx \frac{u_i}{2} \exp\left\{\mathbf{z}_i \boldsymbol{\psi}\right\} \sum_{q=1}^{Q} w_q h_0\left(\frac{u_i(1+t_q)}{2}\right) \exp\left(\eta W_i\left(\frac{u_i(1+t_q)}{2}\right)\right) \\ &\approx \frac{u_i}{2} \exp\left\{\mathbf{z}_i \boldsymbol{\psi}\right\} \sum_{q=1}^{Q} \left\{ w_q \left[\sum_{k=1}^{m_S} \gamma_k^{m_S-1} \frac{f_{Beta}\left(\frac{u_i(1+t_q)}{2T_{max}}; k, m_S - k + 1\right)}{T_{max}}\right] \times \right. \\ &\quad \left. \exp\left(\eta\left(\mathbf{x}_i \boldsymbol{\beta} + (\boldsymbol{\xi}_i^{m_L-1})^\top \mathbf{b}_{m_L-1}\left(\frac{u_i(1+t_q)}{2T_{max}}\right)\right)\right)\right\}. \end{aligned}$$

Note that the only unknown quantity to be estimated on the approximation of the baseline hazard function (Equation (3.7)) is the vector of coefficients $\boldsymbol{\gamma}_{m_S-1}$. Under the Bayesian inference, a prior distribution is attributed for this vector. Here, we highlight that $\gamma_k^{m_S-1}$ represents points of a non-negative function. Therefore, the values that it assumes must be in the domain $(0, \infty)$. In our case, we assumed that $\log(\gamma_k^{m_S-1})$ followed, *a priori*, a Normal distribution. More details about the prior specifications will be discussed ahead in the simulation and real application chapters.

We point out that, although mathematically $h_0(.; m_S)$ converges to the true function $h_0(.)$ when $m_S \to \infty$, one must consider practical aspects such as the number of parameters. In other words, one must choose the degree of BP to provide a good approximation and, at the same time, being careful about the complexity of the model. Moreover, we dealt with examples, both in literature (Wang and Ghosh, 2013) and in practice, in which a relatively low degree was sufficient to provide a good performance.

As an example of a possible value for the degree of the BP, Osman and Ghosh (2012) suggested using m_S as being the smallest integer greater than the square root of n, where n is the sample size (that is, $m_S = \lceil \sqrt{n} \rceil$). Guan (2016) proposed an estimator for the degree of the BP that can be used in general applications. Another alternative is to obtain estimates varying the value of m_S and then choosing the best value by performing a sensitivity analysis. In addition, Wang and Ghosh (2013) proposed criteria for selecting the value of m_S that can be used in the sensitivity analysis. Some works consider the degree as a unknown quantity and proceed with the inferential analysis. We will consider m_S fixed and we propose ways of choosing this value in Section 3.5 (page 51).

For comparison purposes, the structure of the BP can be related with that of splines. However, according to Carnicer and Peña (1993), Farouki (2012) and Osman and Ghosh (2012), the BP has the best approximation between all polynomial approximations, in the sense of preserving the shape of the target function. Moreover, the "knots" in the BP are already specified in opposition to the case of splines. The number of knots is another point to consider; there are works that help with this choice. Nonetheless, by using BP we were able to propose a way of specifying its degree. Finally, Crowther et al. (2012) showed via simulation studies that to model the baseline hazard function with B-splines, specifically, can lead to an underestimation of the linking parameter η when joint modeling longitudinal and survival data. The underestimation depends on the quadrature method used to approximate the cumulative hazard function. We can investigate if this underestimation would occur to the BP. However, we have no such indication or suspicion.

In order to make all these concepts related to the modeling approach via Bernstein Polynomials clearer, next section briefly shows how the BP incorporate data information to provide the approximation for the target functions.

3.3 An Intuition about Bernstein Polynomials

The focus of this section is to give an idea of the role and the importance of each quantity that composes the BP. Here, we aim to shed a light on discussions such as the impact of the choice of T_{max} and practical interpretations of the degree of this polynomial, for example. We will use the same notation as the previous section, but the ideas are also true for the approximation of the longitudinal component.

Consider K a random variable such that, given an observed survival time u_i, it represents the number - say k - of successes in m_S trials. Note that $k \in \{0, 1, \ldots, m_S\}$. Next, the information that u_i brings to the random variable K is the probability of success, which will be u_i/T_{max}, and $T_{max} = \max_{\substack{i\in\{1,2,\ldots,n\}\\ j\in\{1,2,\ldots J_i\}}} (t_{ij}, u_i)$. Thus, we have that $K|(u_i/T_{max}) \sim Binomial(m_S, u_i/T_{max})$. Then,

$$\mathbb{P}\left(K = k|\frac{u_i}{T_{max}}\right) = \binom{m_S}{k}\left(\frac{u_i}{T_{max}}\right)^k\left(1 - \frac{u_i}{T_{max}}\right)^{m_S-k}$$

for $k = 0, 1, \ldots, m_S$; and the relationship between the possible values of K and the observed survival time u_i is given by (for $k = 0, 1, \ldots, m_S - 1$)

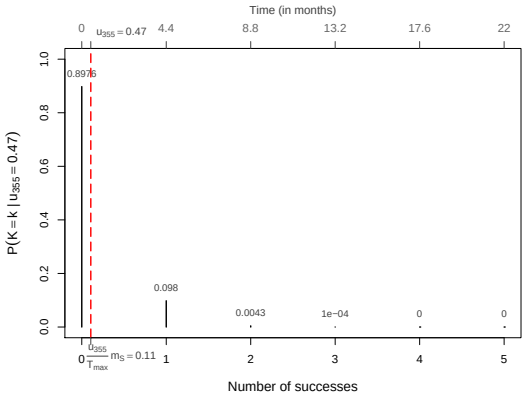
$$\frac{k}{m_s} \leq \frac{u_i}{T_{max}} < \frac{k+1}{m_s} \Leftrightarrow \frac{k}{m_s}T_{max} \leq u_i < \frac{k+1}{m_s}T_{max} \Leftrightarrow k \leq \frac{u_i}{T_{max}}m_s < k+1, \quad (3.9)$$

Figures 3a to 3e are a tentative to clarify this construction. These panels were made based on a data set composed of HIV+ patients that underwent a treatment (this data set will be described with more details in Chapter 5). Here, the information we need from this data set is only the vector of $n = 467$ observed survival times $(u_1, u_2, \ldots, u_{467})$, regardless if they were a failure or a censoring. In this example, we fixed $m_S = 5$ and $T_{max} = 22$. In these panels, the lower x-axis is the number of possible successes $k = 0, 1, \ldots, 5$, and the upper x-axis is the follow-up time varying from $(0, 22]$. These two axis are related according to the equivalences described in Equation (3.9).

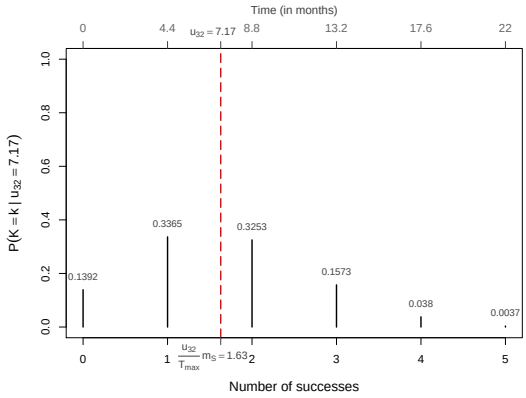
Thus, consider the smallest observed survival time, which was a failure time of $u_{(1)} = u_{355} = 0.47$ month. Figure 3a shows the distribution of K given this failure time. We can see on the upper x-axis the representation of this survival time in the follow-up scale, and in the lower x-axis this same time but in terms of the number of successes. For the minimum survival time, this number was $\frac{u_{355}}{T_{max}}m_S = \frac{0.47}{22}5 = 0.11$. Hence, the figure shows that the probabilities coming from the Binomial distribution are higher for an integer k that is near 0.11; and it diminishes as k gets far from 0.11.

We chose other four observed survival times to reinforce the interpretation of the quantities composing the BP. These times were $u_{32} = 7.17$ (Panel 3b), $u_{133} = 16.70$ (Panel 3d), $u_{283} = 21.40$ (Panel 3e) and $u_{306} = 12.07$ (Panel 3c). By observing all these panels it is clear that conditioning in different observed times redistributes the probabilities of K, which have the role of weights on Bernstein Polynomials. The distribution is such that we have high probabilities around the observed time and it gets smaller as the number of successes gets far from this time. That is, the information that the observed survival data adds to the BP is how to distribute the probabilities that come from a Binomial distribution. So the greater is the degree, the closer the survival times will be to the integer representing the number of successes. In addition, as T_{max} increases, the probability of observing k successes around $\frac{u_i}{T_{max}}m_S$ diminishes.

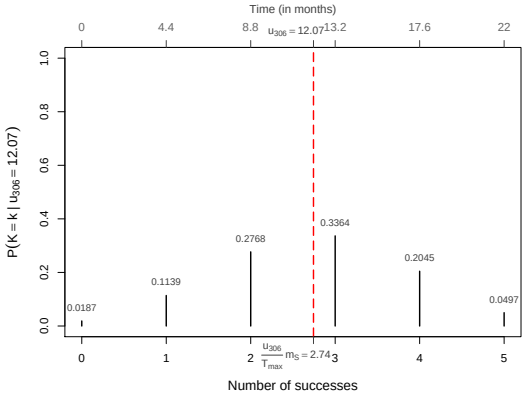
Finally, Figure 3f shows the frequency of the number of successes considering all observed survival times - both failures and censoring. The grey dots are the observed survival times. It is evident that the interval in which we observe more survival times is directly related to the highest frequency of the number of successes.



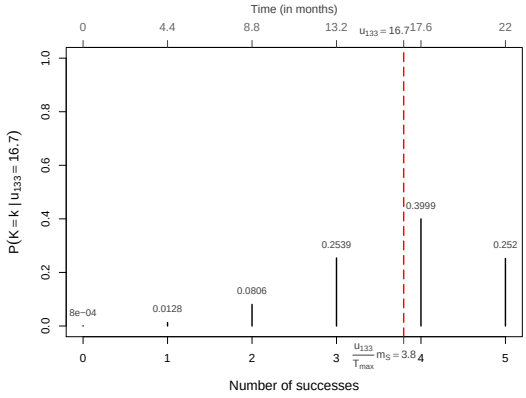
(a) u_{355}.



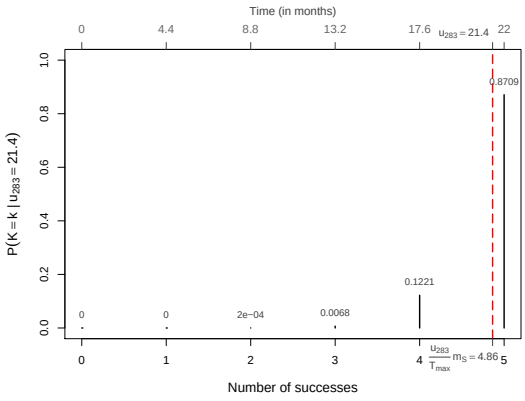
(b) u_{32}.



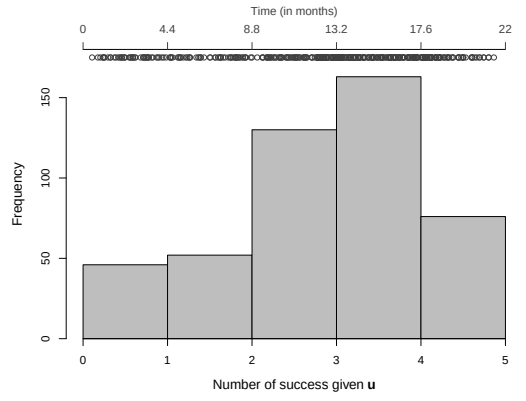
(c) u_{306}.



(d) u_{133}.



(e) u_{283}.



(f) Barplot with all observed survival times.

Figure 3 – Distribution of K given different observed survival times. Each of the panels 3a to 3e take into account only one survival time. Panel 3f is a barplot considering all observed survival times.

Next section discusses about few and important properties of the BP.

3.4 Important Properties of Bernstein Polynomials

In this section we briefly discuss few useful properties of the BP. We will focus on the properties that have a connection with the modeling procedures explored in this work. The introduction of these properties will be very useful to endorse theoretical arguments involving the BP. More properties and details can be found in Lorentz (1986) and Farouki (2012).

The first property refers to the relation between the function representing a straight line and the approximation by BP for this function.

Property 1 (Equivalency with a straight line)**.** *There is a equivalency between BP with an arbitrary degree m and the straight line equation, i. e. $f(t) = at + b$, $a, b \in \mathbb{R}$ and $t \in (0, 1)$. The BP approximation with degree m for this function (see Equation (3.1)) is given by*

$$f(t) \approx \sum_{k=0}^{m} f\left(\frac{k}{m}\right)\binom{m}{k}t^k(1-t)^{m-k} = \sum_{k=0}^{m}\left(a\frac{k}{m}+b\right)\binom{m}{k}t^k(1-t)^{m-k},$$

and it reduces to $at + b$.

Therefore, the mathematical BP approximation for the function $f(t) = at + b$ leads to the very function $f(t)$, exactly. This result will be important for the case where the temporal behavior of the longitudinal variable is expected to present a linear trend. Although we approximate this function by estimating the coefficients of the BP, we assume that results might be slightly more accurate due to this property. It is worth mentioning that we defined the function f on the interval $(0, 1)$ for simplification. In case it is needed, a transformation on the variable t can be applied.

The second property covers both the first and the last coefficients that compose the BP.

Property 2 (First and last Bernstein coefficients)**.** *The first and the last coefficients of a BP with degree m and from a BP with degree $m + r$, $r = 1, 2, \ldots$ represent the same values of the function being approximated. That is,*

$$\xi_0^m \approx f\left(\frac{0}{m}T_{max}\right) = f(0) = f\left(\frac{0}{m+r}T_{max}\right) \approx \xi_0^{m+r}$$

and

$$\xi_m^m \approx f\left(\frac{m}{m}T_{max}\right) = f(T_{max}) = f\left(\frac{m+r}{m+r}T_{max}\right) \approx \xi_{m+r}^{m+r}.$$

In the property above, the role of T_{max} is only to rescale the variable t to the interval $(0, T_{max})$. The practical interpretation of this property is that we expect the estimates of these pair of coefficients to be similar, *i. e.*, $\xi_0^m \approx \xi_0^{m+r}$ and $\xi_m^m \approx \xi_{m+r}^{m+r}$.

The last property is about the degree elevation. This property characterizes the association between the vector of Bernstein coefficients with different degrees. More specifically, given a vector of BP coefficients with degree m it is possible to know exactly - in theory - the vector of coefficients with a degree $m+r$, $r = 1, 2, \ldots$. We will make use of this property to establish a stopping rule for the degree of the BP, which will be discussed on the next section.

Property 3 (Degree elevation)**.** *This property refers to the relationship between the vector of Bernstein coefficients of degree m and $m+r$, $r = 1, 2, \ldots$. It allows us to determine coefficients of a BP with degree $m+r$ given the vector of coefficients of a BP with degree m. This relationship is*

$$\tilde{\xi}_k^{m+r} = \sum_{j=\max(0,k-r)}^{\min(m,k)} \frac{\binom{r}{k-j}\binom{m}{j}}{\binom{m+r}{k}} \xi_k^m, \; k = 1, 2, \ldots, m+r, \tag{3.10}$$

where $\tilde{\xi}_k^{m+r}$ represents the k-th coefficient of a BP with degree $m+r$ obtained via degree elevation and ξ_k^m is the k-th coefficient from a BP with degree m. Note that: (i) the Binomial coefficient terms of the summation in Equation (3.10) come from a $HyperG(m+r, m, k)$ distribution and (ii) if $r = 1$, Equation (3.10) reduces to

$$\tilde{\xi}_k^{m+1} = \frac{k}{m+1}\xi_{k-1}^m + \left(1 - \frac{k}{m+1}\right)\xi_k^m, \; k = 1, 2, \ldots, m \tag{3.11}$$

and $\tilde{\xi}_0^{m+1} = \xi_0^m$ and $\tilde{\xi}_{m+1}^{m+1} = \xi_m^m$.

It means that, theoretically, it is only required a vector of Bernstein coefficients of degree m and then all the other approximations with higher degrees would follow immediately. For this reason, at a first look it may seem dispensable to estimate the vector of coefficients $\boldsymbol{\xi}_{m+r}$, $r = 1, 2, \ldots$ with an entire procedure of estimation. That is, we could simply use Property 3 to instantly obtain this result. However, this behavior of similarity was not observed in practice. At last, note that the results $\tilde{\xi}_0^{m+1} = \xi_0^m$ and $\tilde{\xi}_{m+1}^{m+1} = \xi_m^m$ are in accordance with the second property.

Demonstrations of properties 1 and 3 are on the Appendix A (page 102). In the next section we discuss the guidance we are proposing to help choosing the degree of the BP. This guidance is another contribution of our work.

3.5 Degree Selection

The choice of the degree of the BP represents a great challenge due to the fact that it has an important role in the approximation performance. This characteristic was verified both in literature (Petrone, 1999a; Osman and Ghosh, 2012; Wang and Ghosh, 2013) and in practice. If the degree is too small, the approximated curve may be too simple in the sense of smoothness. Particularly considering the case of survival data, as mentioned in Crowther and Lambert (2013), it is frequent to come across with real data in which the behavior of the hazard function has one or more turning points. Therefore, in these cases the approximation by BP may not encompass such important features of the target function. On the other hand, if this number is too large, there will be an unnecessary large number of parameters. This trait goes against the concept of parsimony and it can lead to computational issues. Nonetheless, the convergence of the BP is only guaranteed when the degree $m \longrightarrow \infty$. This theoretical statement is also true for the accuracy of the estimates for the vector of coefficients, as it is related to the value of the function being approximated (discussed on pages 42 and 44). Then, in a general framework, an "optimal" choice for the degree would be the *minimum* value that manages to approximate well and that accommodates the important features of the target function.

Curtis and Ghosh (2011) fixed the degree and discussed ways of making this choice. Osman and Ghosh (2012) and Wang and Ghosh (2013) also fixed the degree basing their choice of this value on sample sizes. Other works such as Petrone (1999b) and Chang et al. (2005), treated the degree of the BP as an unknown quantity to be estimated. However, considering the complexity of joint modeling longitudinal and survival data, it may be more interesting to fix this quantity with a good strategy. It is then needed a guidance on this choice as well as a sensitivity analysis.

In order to work this obstacle around, we propose a solution that consists on establishing a routine to choose the minimum suitable degree for the BP. This instruction is probabilistic method based on a previous knowledge of a possibly existing turning point on the target function. Here, we mention in anticipation that our method can still be used if one does not know this feature in the function. Our findings point out that the degree of the BP is more associated to how close a change in the behavior of the target function is to the boundaries of the domain, than it is related to sample sizes. Hence, we established a random variable M that represents the minimum degree necessary to capture a change in the target function. We also derived two criteria to serve as stopping rule for the degree of the BP. This stopping rule is based on the degree elevation property (Property 3) seen in Farouki (2012).

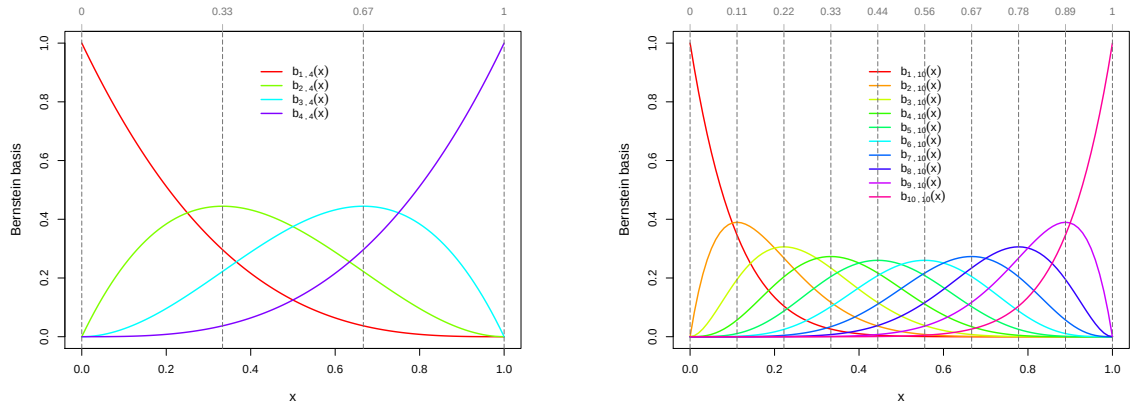
Another point that is worth discussing is about the terminology we use in the next sections. We use expressions such as "turning point" or "change point" in the target function to refer to as a change of increasing/decreasing to decreasing behavior on this

function, *i. e.* local minimum or maximum. It is not associated with the change point literature.

First, note by Equations (3.2) and (3.7) that each coefficient is related to a Bernstein basis $b_{k,m-1}(.)$, $k = 1, 2, \ldots, m$. With these same equations we can recall that the approximation provided by the BP is a linear combination of these two vectors. Since Bernstein basis can be regarded to as weights, it means that each coefficient ξ_k^{m-1} will have more weight on sub-intervals of the domain that takes place around t such that

$$b'_{k,m-1}\left(\frac{t}{T_{max}}\right) = 0 \Leftrightarrow \frac{t}{T_{max}} = \frac{k-1}{m-1} \Leftrightarrow t = \frac{k-1}{m-1}T_{max},\ k = 1, 2, \ldots, m. \tag{3.12}$$

Details of this result can be seen in the Appendix A, page 104. In order to show these maximum values, Figure 2 was revisited in Figure 4. Here, each colored line represent a Bernstein basis, of a BP with degree $m-1$, for $t/T_{max} \in (0, 1)$ and the vertical dotted gray lines represent at which point t each basis reaches its maximum value, according to the relationship in Equation (3.12).



(a) $m = 4$. (b) $m = 10$.

Figure 4 – Illustration of the vector of Bernstein basis, of a BP with degree $m-1$, for $m = 4$ and $m = 10$ and the representation of the time where these functions reach their maximum.

In what follows, one of our interests lies on studying the increasing and decreasing behavior of the target function f being approximated. Thus, we must evaluate the derivative of Bernstein approximation for this function. The BP approximation for f with degree $m-1$ is

$$\begin{aligned} f(t; m-1) &\approx \sum_{k=1}^{m} f\left(\frac{k-1}{m-1}\right)\binom{m-1}{k-1}\left(\frac{t}{T_{max}}\right)^{k-1}\left(1-\frac{t}{T_{max}}\right)^{m-k} \\ &= \sum_{k=1}^{m} \xi_k^{m-1} b_{k,m-1}\left(\frac{t}{T_{max}}\right) \end{aligned} \tag{3.13}$$

and its derivative is

$$\begin{aligned} f'(t;m-1) &= \sum_{k=1}^{m} \xi_k^{m-1} \binom{m-1}{k-1} \left[\frac{(k-1)}{T_{max}} \left(\frac{t}{T_{max}}\right)^{k-2} \left(1-\frac{t}{T_{max}}\right)^{m-k} - \frac{(m-k)}{T_{max}} \left(\frac{t}{T_{max}}\right)^{k-1} \left(1-\frac{t}{T_{max}}\right)^{m-k-1} \right] \\ &= \frac{(m-1)}{T_{max}} \sum_{k=1}^{m-1} (\xi_{k+1}^{m-1} - \xi_k^{m-1}) \binom{m-2}{k-1} \left(\frac{t}{T_{max}}\right)^{k-1} \left(1-\frac{t}{T_{max}}\right)^{m-k-1} \\ &= \frac{(m-1)}{T_{max}} \sum_{k=1}^{m-1} (\xi_{k+1}^{m-1} - \xi_k^{m-1}) b_{k,m-2}\left(\frac{t}{T_{max}}\right). \end{aligned} \tag{3.14}$$

It is well-known that if $f'(t;m-1) < 0$, the curve $f(t;m-1)$ presents a decreasing behavior and if $f'(t;m-1) > 0$, then $f(t;m-1)$ will be an increasing function on t. Also, it is clear that $(m-1)/T_{max} \geq 0$ and $b_{k,m-2}(t/T_{max}) \geq 0$, for $k = 1, 2, \ldots, m-1$. As a conclusion, the only term in Equation (3.14) that controls the increasing/decreasing behavior of this approximation is the vector of coefficients $\boldsymbol{\xi}_{m-1}$. This result is coherent considering that each of the coefficients represents the function f in a specific point of the domain. Therefore, the vector of Bernstein basis does not affect the form of the approximation directly.

If the difference $(\xi_k^{m-1} - \xi_{k-1}^{m-1})$ has an opposite sign than $(\xi_{k+1}^{m-1} - \xi_k^{m-1})$ or, equivalently, if $(\xi_k^{m-1} - \xi_{k-1}^{m-1})(\xi_{k+1}^{m-1} - \xi_k^{m-1}) < 0$, for $k = 2, 3, \ldots, m-2$ we can assure that the approximated function f has a turning point. This behavior is such that it begins at $\xi_{k-1}^{m-1} \approx f\left(\frac{k-2}{m-1} T_{max}\right)$, it happens around $\xi_k^{m-1} \approx f\left(\frac{k-1}{m-1} T_{max}\right)$, and it finally ends at $\xi_{k+1}^{m-1} \approx f\left(\frac{k}{m-1} T_{max}\right)$. Therefore, we need three coefficients to be able to inform this feature. Thus, both k and m impact this result and this method is capable of detecting a change for a time $t \geq 1/(m-1)$ and $t \leq (m-2)/(m-1)$. In other words, we will be able to capture this change only on an interval (a, b) such that $\frac{1}{m-1} < a < b < \frac{m-2}{m-1} \Leftrightarrow m > \max\left(\frac{1}{a} + 1, \frac{2-b}{1-b}\right)$.

From now on, we will focus on the interval $(0, 1)$, because any time scale can be transformed into this range. Thus, consider two random variables U_1 and U_2 defined on $(0, 1)$ that will form an interval $(u_{(1)}, u_{(2)}) \subset (0, 1)$, where $u_{(1)} = \min(U_1, U_2)$ and $u_{(2)} = \max(U_1, U_2)$. Assume a random variable $M = \left\lceil \max\left(\frac{1}{U_{(1)}} + 1, \frac{2-U_{(2)}}{1-U_{(2)}}\right) \right\rceil$, which refers to the minimum degree that is necessary to capture a change in the interval $(u_{(1)}, u_{(2)})$. We have that

$$\begin{aligned} \mathbb{P}(M = m|(U_1, U_2)) &= \mathbb{P}\left(\left\lceil \max\left(\frac{1}{U_{(1)}} + 1, \frac{2-U_{(2)}}{1-U_{(2)}}\right) \right\rceil = m \right) \\ &= \mathbb{P}\left(m - 1 < \max\left(\frac{1}{U_{(1)}} + 1, \frac{2-U_{(2)}}{1-U_{(2)}}\right) \leq m | (U_1, U_2) \right) \\ &= \left[\mathbb{P}\left(U_1 \leq \frac{m-2}{m-1}\right) - \mathbb{P}\left(U_1 < \frac{1}{m-1}\right) \right]^2 - \left[\mathbb{P}\left(U_1 \leq \frac{m-3}{m-2}\right) - \mathbb{P}\left(U_1 < \frac{1}{m-2}\right) \right]^2, \end{aligned} \tag{3.15}$$

for $m = 4, 5, \ldots$; see details in Appendix A (page 104). The cumulative distribution function of M conditioned on (U_1, U_2) is

$$\begin{aligned} F_{M|(U_1,U_2)}(m) &= \sum_{l=4}^{m} \mathbb{P}(M = l|(U_1, U_2)) \\ &= \sum_{l=4}^{m} \left\{ \left[\mathbb{P}\left(U_1 \leq \frac{l-2}{l-1}\right) - \mathbb{P}\left(U_1 < \frac{1}{l-1}\right) \right]^2 - \left[\mathbb{P}\left(U_1 \leq \frac{l-3}{l-2}\right) - \mathbb{P}\left(U_1 < \frac{1}{l-2}\right) \right]^2 \right\} \\ &= \left[\mathbb{P}\left(U_1 \leq \frac{m-2}{m-1}\right) - \mathbb{P}\left(U_1 < \frac{1}{m-1}\right) \right]^2. \end{aligned} \tag{3.16}$$

This result is easily obtained because the summation in Equation (3.16) is a telescope sum. We can see that, since $1/(m-1) \xrightarrow[m\to\infty]{} 0$ and $(m-2)/(m-1) \xrightarrow[m\to\infty]{} 1$, then $F_{M|(U_1,U_2)}(m) \longrightarrow 1$.

Once we have derived the cumulative distribution function of M, it becomes trivial to calculate quantiles of the minimum degree. Table 1 shows some quantiles considering that U_1 and U_2 follow a $Beta(\theta_1, \theta_2)$ distribution. In this case, Equation (3.16) becomes

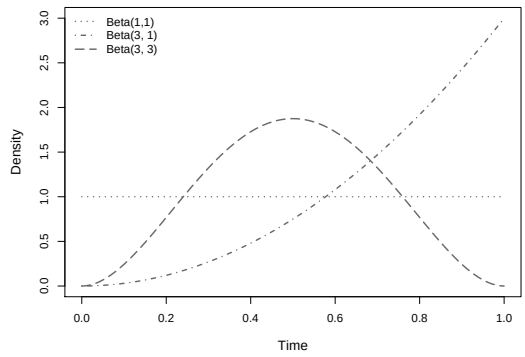
$$F_{M|(U_1,U_2)}(m) = \left[F_{Beta}\left(\frac{m-2}{m-1}\right) - F_{Beta}\left(\frac{1}{m-1}\right) \right]^2.$$

The values of θ_1 and θ_2 were chosen so that we could have high densities near 0 and/or 1 as well as densities concentrated in the middle of this interval. The density functions of the *Beta* distributions with the parameters in the Table 1 can be seen in Figure 5.

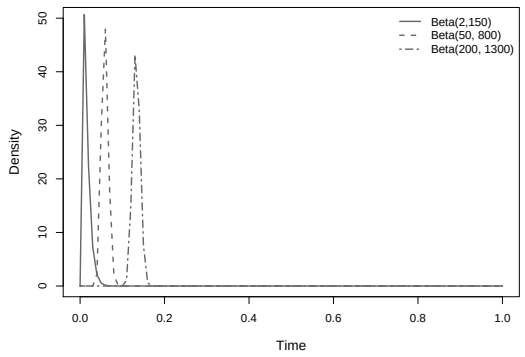
Table 1 – Quantiles of the minimum value of m that is necessary to model a change in the approximated curve on an interval $(U_{(1)}, U_{(2)})$; greatest m such that $\mathbb{P}(M \leq m|(U_1, U_2)) \leq p$.

Distribution of U_1 and U_2	$p = 0.5$	$p = 0.9$	$p = 0.95$	$p = 0.99$	$p = 0.999$
$Beta(1, 1)$	7	39	79	399	3999
$Beta(3, 1)$	10	58	118	598	5998
$Beta(2, 150)$	141	418	618	1453	4709
$Beta(3, 3)$	4	7	9	16	34
$Beta(50, 800)$	19	22	23	25	29
$Beta(200, 1300)$	8	9	9	9	10

As an example, suppose that a change is expected on the interval $(u_{(1)}, u_{(2)})$ such that both U_1 and U_2 follow a $Beta(3, 1)$. This density function is shown in Figure 5a. Then, if $m = 10$ we have a 0.5 probability that the change will be represented in the approximated curve. If we want to be more precise about this result, we should consider $m \geq 118$, which is a very high degree.



(a) Examples of densities spread on the entire interval.



(b) Examples that have densities more concentrated at the lower boundary.

Figure 5 – Density functions of the Beta distributions used in the examples described in Table 1.

The first example in Table 1 is an $Uniform(0,1)$ distribution; in this case we would have no prior information about where the function being approximated may change its behavior, in what concerns the increasing/decreasing course of this function. We can see that $\mathbb{P}(M \leq 39|(U_1, U_2)) \approx 0.9$. In what follows, the next two examples $Beta(3,1)$ and $Beta(2,150)$ admit high density on the lower boundary. So, it is required a large minimum degree in order for the BP to be able to approximate this behavior. On the other hand, if the focus relies far from 0 and 1, m can be considerably lower, as in the last three rows of Table 1.

In practice, one can follow the subsequent step-by-step in order to choose the minimal degree m:

Step 1: In which part of the domain the curve will probably change behavior? If the domain is other than $(0,1)$, then proceed to the next step with a transformation on the scale;

Step 2: Find (θ_1, θ_2) such that $Beta(\theta_1, \theta_2)$ has high density on the region above;

Step 3: Find quantiles of minimum m by solving $\mathbb{P}(M \leq m|(U_1, U_2)) = p \Leftrightarrow F_{Beta}\left(\frac{m-2}{m-1}; \theta_1, \theta_2\right) - F_{Beta}\left(\frac{1}{m-1}; \theta_1, \theta_2\right) - \sqrt{p} = 0$. The command `uniroot` in `R` can be used to solve this equation.

The instructions above are a conclusion that leads to a probabilistic view of a suitable minimum degree for the BP. We highlight that all the discussion and results involving this minimum value is a contribution of our work. The final results of this chapter focus on completing the guidance on choosing the degree by discussing a reasonable maximum value for m.

Degree elevation to achieve optimal m

The stopping rule we will discuss in this section was based on the degree elevation property, seen in Farouki (2012). Regarding this property, it is intuitive to expected that estimates of Bernstein coefficients obtained either via a direct way or using degree elevation should be similar. That is, $|\xi_k^m - \tilde{\xi}_k^m| \approx 0$ for all k. However, in practice, we observed that when m is small there can be a significant difference between estimates based on $\boldsymbol{\xi}_{m-1} = (\xi_1^{m-1}, \xi_2^{m-1}, \ldots, \xi_m^{m-1})^\top$ and $\boldsymbol{\xi}_m = (\xi_1^m, \xi_2^m \ldots, \xi_{m+1}^m)^\top$. Consequently, this discrepancy leads to a difference between $\tilde{\boldsymbol{\xi}}_m$ and $\boldsymbol{\xi}_m$. Nevertheless, when m is large enough, the estimated curve stabilizes and the difference between these two vectors of coefficients gets considerably small. Based on this discussion, we came up with two criteria that, along with regular model comparison measures, can help us to provide an optimal degree for the BP. These two criteria is another contribution of this thesis.

Criterion 1 - difference between coefficients

Under a Bayesian framework, let $D_{m-1}^{(s)} = \dfrac{1}{m}\displaystyle\sum_{k=1}^{m} |\xi_k^{m-1(s)} - \tilde{\xi}_k^{m-1(s)}|$, for $m \geq 5$ and $s = 1, 2, \ldots, S$. The quantity $D_{m-1}^{(s)}$ represent the difference between coefficients obtained via degree elevation ($\tilde{\xi}_k^{m-1}$) and direct estimation (ξ_k^{m-1}) based on the s-th posterior sampled values. If m is small, then it is expected that Median($\mathbf{D}_{m-1}$) > Median($\mathbf{D}_m$), where $\mathbf{D}_m = (D_m^{(1)}, D_m^{(2)}, \ldots, D_m^{(S)})^\top$. However, if m is considered to be large enough there will be no significant difference between these two quantities, meaning that we have reached an optimal degree.

Hence, we will test the hypothesis H_0 : Median($\mathbf{D}_{m-1}$) = Median($\mathbf{D}_m$) vs. H_1 : Median($\mathbf{D}_{m-1}$) > Median($\mathbf{D}_m$) and we will increase m unity by unity until we no longer reject the null hypothesis. Thus, the optimal degree is given by $m_{opt} = \min\{m > 4 :$ p-value $> 0.1\} + 1$. Here, $m \geq 5$ because our proposed method for the minimum degree is available for $m \geq 4$. Then, we will start the comparison of the direct estimation for $m = 5$ with the degree elevation method based on the posterior sample of a BP with $m = 4$.

Criterion 2 - difference between estimated curves

The estimation for the function of interest, as defined in Equation (3.13), can be rewritten as $f(t; m-1) = (\boldsymbol{\xi}_{m-1})^\top \mathbf{b}_{m-1}(t/T_{max})$, where $\mathbf{b}_{m-1}(t/T_{max})$ is the vector of Bernstein basis and $\boldsymbol{\xi}_{m-1}$ is the vector of coefficients. The second criterion is based on the distance between estimated curves based on both $\boldsymbol{\xi}_{m-1}$ and $\tilde{\boldsymbol{\xi}}_{m-1}$. This distance will be defined as

$$
\begin{aligned}
D_{m-1} &= \int_0^1 (f(t;m-1) - \tilde{f}(t;m-1))^2 d(t/T_{max}) \\
&= \int_0^1 \left((\boldsymbol{\xi}_{m-1})^\top \mathbf{b}_{m-1}\left(\frac{t}{T_{max}}\right) - (\tilde{\boldsymbol{\xi}}_{m-1})^\top \mathbf{b}_{m-1}\left(\frac{t}{T_{max}}\right)\right)^2 d(t/T_{max}) \\
&= \int_0^1 \left[\sum_{k=1}^{m}\sum_{l=1}^{m} (\xi_k^{m-1} - \tilde{\xi}_k^{m-1})(\xi_l^{m-1} - \tilde{\xi}_l^{m-1}) b_{k,m-1}\left(\frac{t}{T_{max}}\right) b_{l,m-1}\left(\frac{t}{T_{max}}\right)\right] d(t/T_{max}) \\
&= \sum_{k=1}^{m}\sum_{l=1}^{m} (\xi_k^{m-1} - \tilde{\xi}_k^{m-1})(\xi_l^{m-1} - \tilde{\xi}_l^{m-1}) \int_0^1 \left[b_{k,m-1}\left(\frac{t}{T_{max}}\right) b_{l,m-1}\left(\frac{t}{T_{max}}\right)\right] d(t/T_{max}) \\
&= \sum_{k=1}^{m}\sum_{l=1}^{m} (\xi_k^{m-1} - \tilde{\xi}_k^{m-1})(\xi_l^{m-1} - \tilde{\xi}_l^{m-1}) \binom{m-1}{k-1}\binom{m-1}{l-1} \frac{\Gamma(k+l-1)\Gamma(2m-k-l+1)}{\Gamma(2m)} \\
&= \sum_{k=1}^{m}\sum_{l=1}^{m} (\xi_k^{m-1} - \tilde{\xi}_k^{m-1})(\xi_l^{m-1} - \tilde{\xi}_l^{m-1}) \frac{1}{(2m-1)} HyperG(k-1; 2m-2, k+l-2, m-1) \\
&= (\boldsymbol{\xi}_{m-1} - \tilde{\boldsymbol{\xi}}_{m-1})^\top \mathrm{A} (\boldsymbol{\xi}_{m-1} - \tilde{\boldsymbol{\xi}}_{m-1}),
\end{aligned} \tag{3.17}
$$

where A is an $m \times m$ matrix such that $\mathrm{A} = [a_{kl}]$, for $k = 1, 2, \ldots, m$, $l = 1, 2, \ldots, m$ with $a_{kl} = \dfrac{1}{(2m-1)} HyperG(k-1; 2m-2, k+l-2, m-1)$. At last, $HyperG(k-1; 2m-2, k+l-2, m-1)$ is the probability of $k-1$ successes from an *Hypergeometric* distribution with parameters $(2m-2, k+l-2, m-1)$. See this result with details in the Appendix A, page 105.

Here, we are interested in testing H_0 : Median$(\mathbf{D}_{m-1})$ = Median$(\mathbf{D}_m)$ vs. H_1 : Median$(\mathbf{D}_{m-1})$ > Median$(\mathbf{D}_m)$. Similarly to the first criterion, to reject the null hypothesis means that m is still low and, as a result, a better estimated curve can be obtained by increasing the degree by a unity. The optimal value of m is $m_{opt} = \min\{m > 4 : \text{p-value} > 0.1\} + 1$.

The difference between the proposed criteria is that the first one considers only BP coefficients, which evaluate the approximated curve difference in m points. The second one takes into account the entire approximation difference. In order to test these hypothesis, we can use Sign Test and/or Wilcoxon Rank Test; see more about these tests in Gibbons and Chakraborti (2003). In addition, both tests are available in `R`. The Sign test can be accessed using the package `BSDA` (Arnholt and Evans, 2017) and the Wilcoxon test at the `R` basis.

In this chapter we have introduced the Bernstein Polynomials and we described how to use this tool to approximate target functions. In our case, these functions were the baseline hazard function / cumulative baseline hazard function and the time-varying part of the longitudinal component. We also gave an intuition of how this mechanism uses data information in the approximation procedure. Moreover, we discussed properties and we proposed methods to choose an optimal degree.

We remind that our main contribution is to jointly model longitudinal and survival data using Bernstein Polynomials in both sub-models. Besides that, the method that we

proposed for choosing for the minimum and optimal values of the degree is a novelty. Next chapter focuses on simulation studies.

4 Simulation Studies

In this thesis we conducted two simulation studies. The main goal of the first one was to verify the performance of the two proposed degree selection methods. We aimed at investigating if the theoretical results that we derived could be put into practice. In the second simulation study, our purpose was to (i) check the performance of the proposed model, and (ii) compare our proposal to other models available in the literature. Description, discussion and results of these studies will be covered in this chapter.

The comparison measures we adopted in this work were the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002), the Logarithm of the Pseudo-Marginal Likelihood (LPML) (Gelman et al., 2003) and the Watanabe-Akaike Information Criterion (WAIC) (Watanabe, 2010; Vehtari and Gelman, 2014). The last two measures will be shown multiplied by -2 in order to be in the same scale as the DIC. These measures tell which model is the best one within all the fitted models. So, in this comparison, *the lower* these measures are, *the better* the model is. Details about these criteria can be seen in the Appendix B (page 107).

Analyses were performed using the software `R` (R Core Team, 2019) and the platform `Stan` (Carpenter et al., 2017; Stan Development Team, 2018). In order to connect `R` and `Stan`, the package `RStan` (Stan Development Team, 2019) was used.

4.1 Evaluation of the degree selection methods

Suppose that the true curve of a certain phenomena evolves according to a function $f(t) = 10 + 10\sin(2\pi t)$, $t \in (0, 1)$. The random variable referring to the observed values is $Y_i(t_{ij})$ for subjects $i = 1, 2, \ldots, n$ and measurements $j = 1, 2, \ldots, J_i$. Each observation is equal to the "true" and unobserved value $f(t_{ij})$ plus a measurement error:

$$Y_i(t_{ij}) = W_i(t_{ij}) + \epsilon(t_{ij}) = f(t_{ij}) + \epsilon(t_{ij}). \tag{4.1}$$

Here, $W_i(t_{ij})$ represents the true value of the longitudinal variable. This true value behaves according to the function f. Next, $\epsilon(t_{ij})$ is an independent and identically normally distributed random error with mean 0 and variance σ^2_ϵ.

We generated 100 data sets with sample size $n = 50$ each. The number of measurements were uniformly distributed within the set $\{3, 4, \ldots, 10\}$. The first measurement time was always 0 (*i. e.*, $t_{i1} = 0 \ \forall \ i$) and the measurement times for $i = 1, 2, \ldots, n$ and $j = 2, 3, \ldots, J_i$ were sampled from a $Beta(1, 3)$ distribution. This distribution

has high density at the beginning of the interval $(0, 1)$, representing the idea that there are more measurements at initial times and fewer ones at final times. Thus, $\forall\ i$ we sampled $\mathbf{W}_i = (W_i(t_{i1}), W_i(t_{i2}), \ldots, W_i(t_{iJ_i}))^\top \sim Normal_{J_I}(\mathbf{f}_i, \Sigma_{J_i})$, where $\mathbf{f}_i = (f(t_{i1}), f_i(t_{i2}), \ldots, f_i(t_{iJ_i}))^\top$, $\Sigma_{J_i} = [\sigma_{jj'}]$, $\sigma_{jj'} = 6$ if $j \neq j'$ and $\sigma_{jj'} = 3^2$ if $j = j'$. At last, we used Equation (4.1) to obtain the observed values $\mathbf{y}_i = (y_i(t_{i1}), y_i(t_{i2}), \ldots, y_i(t_{iJ_i}))$, $\forall\ i$, with $\sigma_\epsilon = 1.5$. We highlight that the structure of these data being in a longitudinal configuration was merely for convenience, and it does not favor our degree selection method.

The true curve $f(t)$ is illustrated on Figure 6. We can see that it changes behavior in two points, at $t = 1/4$ and $t = 3/4$. Therefore, considering that this information is known by an expert, the next step is to translate this knowledge to Beta distributions with high densities in intervals including these two points (see step by step in page 55). Then, we chose $Beta(11, 34)$, $Beta(117, 351)$ and $Beta(1172, 3515)$. The expected values for these distributions are 0.2444, 0.2500 and 0.2501 and the variances are 0.0040, 0.0004 and 0.00004.

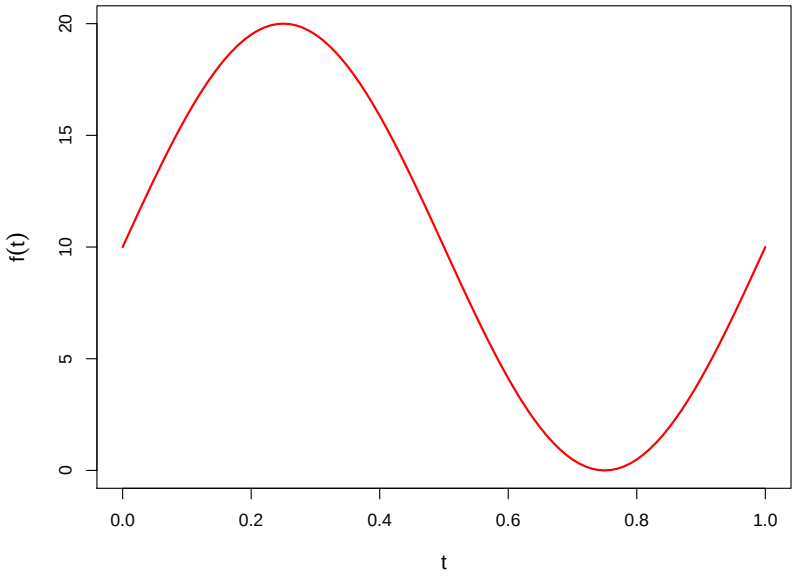


Figure 6 – Graph of the function $f(t) = 10 + 10\sin(2\pi t)$, $t \in (0, 1)$.

In this example both turning points are equally spaced in relation to the boundaries. So, in this specific case, we can consider only one of them, because the indicated minimum value will be exactly the same. We chose the first one, $t = 1/4$. In a case where the points in which a change of behavior is expected are not equally spaced, one can apply the step by step for each one of them. Then, the final minimum degree would be the maximum of the indicated values. Another strategy would be to choose the point that is closer to the boundaries.

The representation of these densities along with the true mean curve can be observed at Appendix B, Figure 24 (page 109). Tables 2 and 3 show the probability function and

the cumulative distribution function of the random variable M given the three chosen Beta distributions for U_1 and U_2. We can see that the highest probability for all three cases is when $m = 6$. In addition, $\mathbb{P}(M \geq 6|(U_1, U_2)) \geq 0.5$ in all examples. Thus, a suitable minimum value for m in this example is 6. We will use the proposed methods to point out the optimal degree for the BP that best approximates $f(t)$.

Table 2 – Probability function of M given different distributions for U_1 and U_2, *i. e.*, $\mathbb{P}(M = m|(U_1, U_2)))$.

Distribution of	m											
U_1 and U_2	4	5	6	7	8	9	10	11	12	13	14	...
$Beta(11, 34)$	0.0078	0.1878	0.3612	0.2463	0.1148	0.0477	0.0195	0.0082	0.0035	0.0016	0.0008	...
$Beta(117, 351)$	0.0000	0.2430	0.7480	0.0090	0.0000	0.0000	0.000	0.0000	0.0000	0.0000	0.0000	...
$Beta(1172, 3515)$	0.0000	0.2511	0.7489	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	...

Table 3 – Cumulative distribution function of M given different distributions for U_1 and U_2, *i. e.*, $\mathbb{P}(M \leq m|(U_1, U_2)))$.

Distribution of	m											
U_1 and U_2	4	5	6	7	8	9	10	11	12	13	14	...
$Beta(11, 34)$	0.0078	0.1957	0.5569	0.8032	0.9180	0.9656	0.9851	0.9933	0.9969	0.9985	0.9992	...
$Beta(117, 351)$	0.0000	0.2430	0.9909	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	...
$Beta(1172, 3515)$	0.0000	0.2511	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	...

For each data set, we fitted the model in Equation (3.2) with orders in the set $\{4, 5, \ldots, 16\}$. For the MCMC specification, we set a burn-in of 50,000, a lag of 20 and we saved 5,000 posterior values. The prior distributions were weakly informative: $\boldsymbol{\mu}_\xi \sim Normal_{m_L}(\mathbf{0}_{m_L}, (10)^2\mathbb{I}_{m_L})$, $\Sigma_\xi^{-1} \sim Wishart(m_L + 2, (1/m_L)\mathbb{I}_{m_L})$, and $1/\sigma_\epsilon^2 \sim Gamma(0.01, 0.01)$. Here, $\mathbf{0}_{m_L}$ represents a vector of length m_L in which each component is equal to 0, and $\mathbb{I}_{m_L}$ stands for an $m_L \times m_L$ identity matrix.

Our aims were (i) to compare the best degree for the BP suggested by both proposed criteria; and (ii) given the optimal degree, to indicate when changes in the target curve occur. Moreover, since we know the true curve function $f(t)$, we can use both proposed methods to obtain the optimal degree by comparing the estimates $\boldsymbol{\xi}_{m-1} = (\xi_1^{m-1}, \xi_2^{m-1}, \ldots, \xi_m^{m-1})^\top$ and the true values that this vector of parameters represents, which are $f((l-1)/(m-1))$, for $l = 1, 2, \ldots, m$.

Table 4 shows results regarding the optimal degree for the BP. The contents in this table are the frequencies in which each degree was selected as the best one. This selection varied according to the criterion (1 or 2, see Section 3.5 on page 56) and the non-parametric tests (Sign or Wilcoxon). The left side of this table is based on Criterion 1, *i. e.*, difference between coefficients. We compared the difference between *estimated BP coefficients versus BP coefficients obtained via degree elevation* as well as the difference

between *estimated BP coefficients versus true BP coefficients*. In all these cases, both Sign and Wilcoxon tests indicated that $m_{opt} = 10$. Nonetheless, when we compare the direct estimation against points of the true curve (columns 3 and 4), there was not a clear difference between $m = 10$ and $m = 11$. Since in the real data scenario we do not know the true curve, we will choose the best degree as $m_{opt} = 10$. Notwithstanding, merely for comparison purposes, we will also discuss results when $m_{opt} = 11$.

Table 4 – Optimal degree for BP (m_{opt}) based on proposed criteria.

	Difference between coefficients				Difference between mean curves			
	degree elevation		true		degree elevation		true	
m	Sign	Wilcoxon	Sign	Wilcoxon	Sign	Wilcoxon	Sign	Wilcoxon
5	0	0	0	0	0	0	0	0
6	0	0	0	1	**65**	**60**	0	0
7	0	0	0	0	12	7	2	2
8	14	11	8	6	4	5	3	3
9	16	17	7	8	11	10	2	2
10	**47**	**45**	**35**	**36**	4	5	15	15
11	20	24	34	35	3	8	**18**	**17**
12	3	3	14	13	1	4	15	13
13	0	0	0	0	0	1	17	14
14	0	0	2	1	0	0	7	10
15	0	0	0	0	0	0	8	8
16	0	0	0	0	0	0	4	4
≥ 17	0	0	0	0	0	0	9	12

In turn, the right side of Table 4 shows results for the same comparisons above based on the differences between the entire mean curves (Criterion 2). Thus, we have differences between *direct estimated mean curve versus mean curve via degree elevation* as well as the difference between *direct estimated mean curve versus true mean curve*. According to Criterion 2, the best degree is $m_{opt} = 6$; on the other hand, if we compare estimated results to the true mean curve, $m_{opt} = 11$. Using the same arguments above, this method selects the best degree $m_{opt} = 6$ but it is also worth discussing results when $m_{opt} = 11$.

Based on the previous discussion, we evaluated the probability of change in the mean curve for $m = 6, 10$ and 11, see Figure 7. On the left panel of this figure (Figures 7a, 7c and 7e) we can observe boxplots based on 100 posterior probabilities of a change in the mean curve. These probabilities are defined as $\mathbb{P}\left((\mu_{\xi_l}^{m-1} - \mu_{\xi_{l-1}}^{m-1})(\mu_{\xi_{l+1}}^{m-1} - \mu_{\xi_l}^{m-1}) < 0 | \text{Data}\right)$, for $l = 2, 3, \ldots, m-2$ and they can be calculated as

(i) fix one data set and compute the difference between adjacent coefficients $(\mu_{\xi_l}^{m-1} - \mu_{\xi_{l-1}}^{m-1})$, for $l = 2, 3, \ldots, m$, using the posterior sample. It leads to a $5000 \times (m-1)$

matrix. We can call this matrix S_1;

(ii) verify the sign of each element of the matrix S_1. Let this matrix with the sign representations be called S_2;

(iii) build matrix S_3 with dimensions $(5000 \times (m-2))$. In order to fill matrix S_3 we will compare the signs of matrix S_2. Then, for the s-th posterior sample if the signs of $(\mu_{\xi_l}^{m-1} - \mu_{\xi_{l-1}}^{m-1})$ and $(\mu_{\xi_{l+1}}^{m-1} - \mu_{\xi_l}^{m-1})$ are equal, the (s,l)-th element is equal to zero, for $s = 1, 2, \ldots, 5000$ and $l = 1, 2, \ldots, m-2$. Otherwise, it is equal to 1;

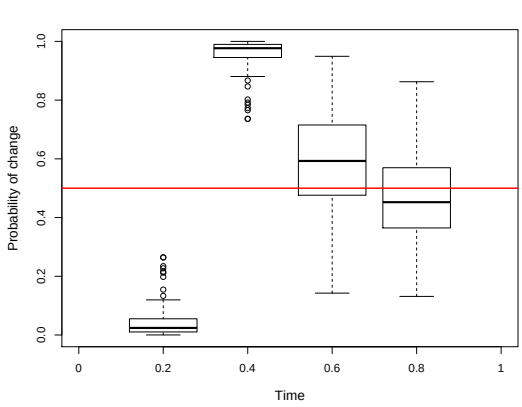
(iv) compute a vector of the posterior probability of a change by checking how many values in each column of matrix S_3 were equal to 1. This number is divided by the posterior sample size 5000. This vector is composed of $m-2$ components;

(v) repeat the procedures above for the results of all generated data sets.

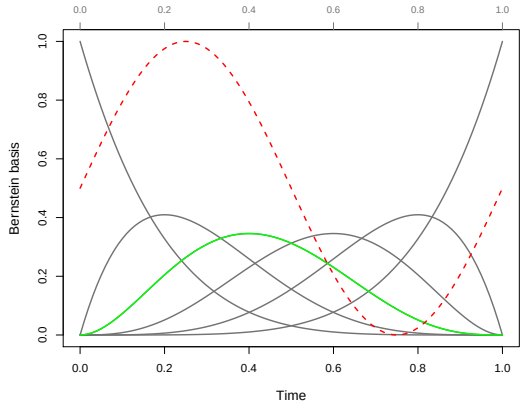
The procedures above led us to the results shown in the boxplots of the left panel of Figure 7. Each value of the vector in item (iv) refers to the possibility of a change in one of the $m-2$ time points. The red line in these figures represent a cutting value of 0.5. In addition, if the model captures a change in the target function, it starts in $t = (l-2)/(m-1)$, happens around $t = (l-1)/(m-1)$ and finishes on $t = l/(m-1)$, $l = 2, 3, \ldots, m-1$. It means that it is necessary the information of three coefficients in order to capture a change; therefore, there will be $m-2$ boxplots in each figure of the left panels.

The right panel in Figure 7 shows each of the m Bernstein basis. The red dotted line is the true curve f after it was reescaled to the $(0,1)$ interval. We reescaled this function so that it could fit in the figures of the basis and, therefore, facilitate the comparisons. The colored straight lines represent the bases related to the coefficients that, in turn, represent at each time point the mean curve is expected to change its behavior. We determined that it is *likely* that a change will occur if the median of the probabilities for the data sets was equal or above 0.6.

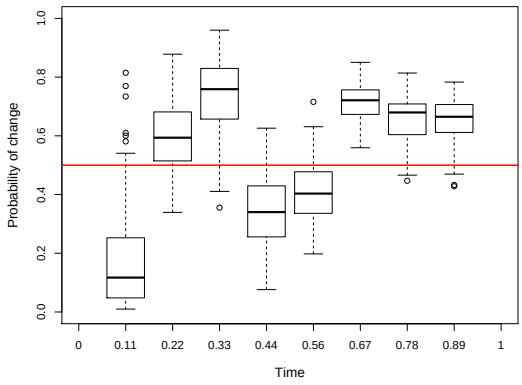
Consider $m = 6$ (Figures 7a and 7b); this value was selected as the optimal degree according to Criterion 2. We can see that the posterior probabilities of change around $t = 0.2$ are all below 0.5. That is, results show that is *extremely* ***unlikely*** that a change starts in $t = 0$, happens around $t = 0.2$ and ends until $t = 0.4$. The second boxplot is entirely above the straight line 0.5; thus, it is *extremely* ***likely*** that a change will occur around $t = 0.4$. In turn, the median of the third boxplot is above 0.5, therefore we can consider that there will be another change around $t = 0.6$ and, at last, the fourth boxplot indicates that no change is likely to occur in $t = 0.8$. Figure 7b shows all $m = 6$ Bernstein basis and the colored ones are those related to the coefficients that detected a change in the mean curve. We can see that the BP detects the changes even with a small degree.



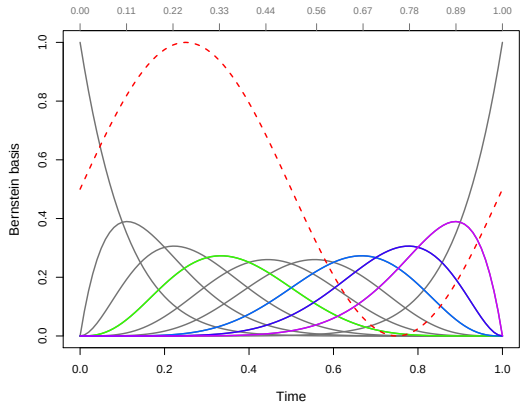
(a) $m = 6$.



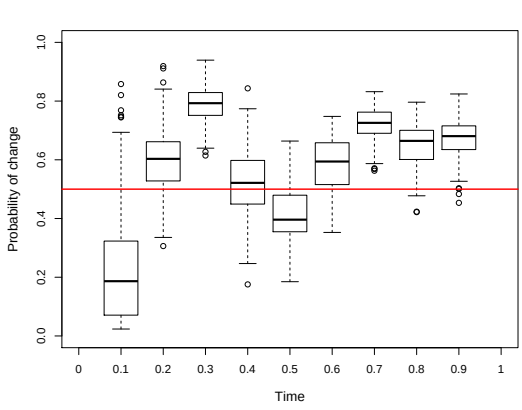
(b) $m = 6$.



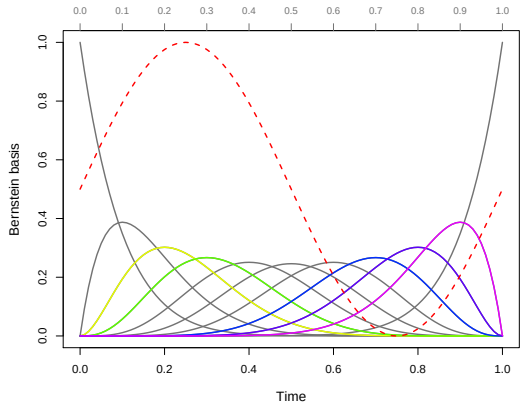
(c) $m = 10$.



(d) $m = 10$.



(e) $m = 11$.



(f) $m = 11$.

Figure 7 – Posterior probability of a change point in the mean curve (left panel) and basis related to these changes (right panel), for $m = 6, 10$ and 11.

However, this degree was too low to allow the estimation to detect this change at a precise time point.

The same analysis was done considering $m = 10$. We can see in Figure 7c that the results indicate that changes occur in the intervals $t \in (0.11, 0.44)$ and $t \in (0.56, 1)$. The highest probabilities are on $t = 0.33$ and $t = 0.67$. An interesting aspect to point out

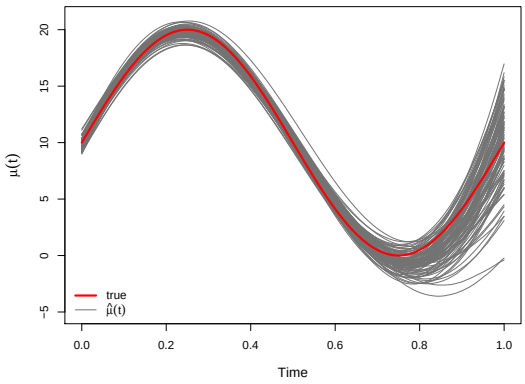
is that, since there were fewer observations at final times, there are more uncertainty in these estimates. The implication of this characteristic is that it leads to a wider interval pointing to where the change might occur. In Figure 7d we can see that there is one basis related to the first change and three basis related to the second one. We have more basis related to the second change precisely due to less data at this part.

In Figure 7e we can see results indicating that a change will occur around $t \in (0.1, 0.5)$ with higher probability centered on $t = 0.3$. Another change is detected on $t \in (0.5, 1)$, being most likely to happen around $t = 0.7$. Figure 7f shows the basis functions that are related to the changes. Similar to what we discussed for $m = 10$, here we also have more uncertainty related to the second change due to fewer data.

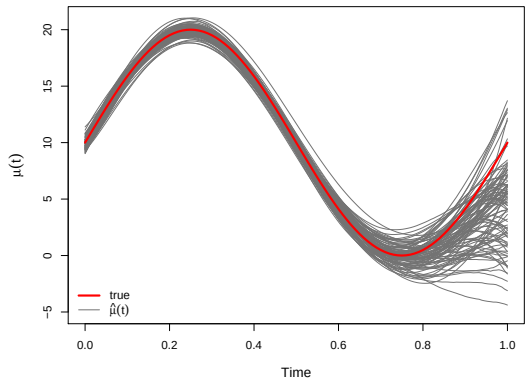
Figure 8 shows the median of the estimated posterior mean curves for each of the 100 data sets (*i. e.*, median of estimated mean function based on all 5000 posterior values) and $m = \{6, 10, 11\}$. The true mean curve is the straight red line. In a general aspect, performance based on the three selected degrees presented similar results. Then, we must consider the trade-off between a higher degree (in this case $m = 10$ or $m = 11$) and the concept of parsimony. A higher degree leads to a more accurate indication of time when the change happens, due to the fact that the proposed method detects changes around a time t such that $t = (l - 1)/(m - 1)$. However, regarding the overall mean curve the lower degree ($m = 6$) is enough to provide good estimates. In addition, the variability of estimates at the end of follow up is larger due to fewer data.

This simulation study indicates that the two methods we propose for the degree selection presented adequate results. Comparing the outcomes, both BP with degrees $m = 6$ and $m = 10$ were able to approximate well the target function. On one hand, estimates based on the lower degree presented lower variability even at the final times - this region is where we had less data. On the other hand, the greater is the degree of the BP, the more accurate it will point out the time points in which the changes occur. That is, when $m = 6$ the possible times points evaluated were $t \in \{0.2, 0.4, 0.6, 0.8\}$; in turn, if $m = 10$, this approximation has the set $t \in \{0.1, 0.2, \ldots, 0.8, 0.9\}$ to point out the changes. Thus, in the cases where the methods we proposed are not in accordance with each other, one should prioritize or (i) the final result of the estimation of the target function; or (ii) the will of being able to tell when the approximate curve will change behavior. In the first case, one should choose the lower indicated degree; on the other hand, if the main goal is (ii) then the choice of the optimal degree should be the maximum value indicated by our proposed criteria.

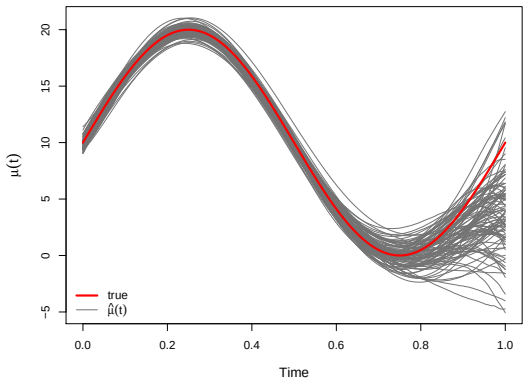
The next sections focuses on the second simulation study. Our main goal with this study was to discuss the performance of the proposed model, *i. e.*, joint modeling longitudinal and survival data using BP in both sub-models.



(a) $m = 6$.



(b) $m = 10$.



(c) $m = 11$.

Figure 8 – Overall median of the posterior mean curves along with the true curve, for $m = 6, 10$ and 11.

4.2 Evaluating the proposed modeling approach

It was described on Section 2.3 that joint models are composed of two sub-models. At the basics, these sub-models consist of a NLME for the longitudinal part and a proportional hazards framework for modeling survival times.

Focusing on the survival sub-model, Bender et al. (2005) described a way of generating survival data based on the parametric proportional hazard model. This method can be applied for any distribution using the baseline cumulative hazard function $H_0(.)$ and its inverse $H_0^{-1}(.)$; therefore, it includes usual and non usual distributions for the failure times. Following this method, Austin (2012) showed how to simulate survival data from a proportional hazards model with a time-dependent covariate. This paper considers the case of continuous and binary time-dependent variables; the latter can present one or more changes through time. The author also derived calculations to obtain closed-form expressions to generate survival times if we assume that they follow an Exponential or a Gompertz distribution.

Despite of these data generation methods for survival times, Bender et al. (2005) and Crowther et al. (2012) claim that most common distributions for these times such as Exponential, Gompertz, and Weibull may be restrictive and, therefore, fail to represent practical situations. One of these situations, for example, is the hazard function presenting one or more turning points. According to Crowther and Lambert (2013), it is frequent to come across real data with such peculiarity. In this context, Crowther et al. (2012) overcame this challenge by generating survival times assuming that they follow a mixture of Weibull distributions. By using this mixture, it is possible to obtain flexible hazard curves, including the mentioned situation of a turning point. After defining this distribution for the survival times, the authors followed the method proposed by Bender et al. (2005) to generate survival times. Crowther and Lambert (2013), among other characteristics such as time dependent effect, showed how to simulate joint longitudinal and survival data with a flexible baseline hazard function.

The scenario of the simulation that we will explore in this section consists of a NLME Model for the longitudinal sub-model and a Gompertz distribution for the survival times. These joint data were generated following the step by step found in Crowther and Lambert (2013) and the expressions obtained by Austin (2012). This specific configuration has the advantage that the cumulative hazard function can be obtained in a closed-form, even considering the time-varying covariate. In this case, although the form of the baseline hazard function is necessarily a monotone function of time and, due to this reason, it will not present a change point, it is not as trivial as a constant function.

Following the methods discussed above, the procedure to generate joint longitudinal and survival data is:

1. obtain the cumulative hazard function $H(.)$;
2. find its inverse function $H^{-1}(.)$;
3. sample survival times following Bender et al. (2005);
4. define event and censored times considering a hypothetical maximum follow-up time (type I scheme of censorship);
5. define measurement times based on the observed survival times;
6. generate longitudinal data.

A relevant observation is that there can be variations in the scheme described above. More details about them can be seen in Crowther and Lambert (2013).

Regarding the longitudinal sub-model, it is expected that the approximation by the BP presents results especially similar to the true model. This expectation is due to the

fact that the BP can approximate well smooth curves, but mainly motivated by the result of Property 1 (Section 3.4, page 49). The formulation of both longitudinal and survival sub-models are given by Equations (4.2) and (4.3), respectively.

$$\begin{aligned} Y_i(t_{ij}) &= W_i(t_{ij}) + \epsilon_i(t_{ij}) \ , \ i = 1, 2 \ldots n \ , \ j = 1, 2, \ldots, J_i \\ &= \mathbf{x}_i\boldsymbol{\beta} + b_{0i} + b_{1i}t_{ij} + \epsilon_i(t_{ij}) \\ &= x_{i1}\beta_1 + x_{i2}\beta_2 + b_{0i} + b_{1i}t_{ij} + \epsilon_i(t_{ij}), \end{aligned} \tag{4.2}$$

$$\begin{aligned} h_i(u_i) &= h_0(u_i)\exp\{\eta W_i(u_i) + \mathbf{z}_i\boldsymbol{\psi}\} \\ &= \lambda e^{\alpha u_i}\exp\{\eta\left[x_{i1}\beta_1 + x_{i2}\beta_2 + b_{0i} + b_{1i}u_i\right] + z_{i1}\psi_1 + z_{i2}\psi_2\}, \end{aligned} \tag{4.3}$$

where $\epsilon_i(t_{ij}) \sim Normal(0, \sigma^2_\epsilon = 1.5^2)$ is the i. i. d. measurement error. We considered a sample size of $n = 400$. Since the number of measurements for each subject depends on the generated survival times, this quantity was also random. There were two covariates: one continuous, $\mathbf{x}_1 = (x_{11}, x_{12}, \ldots, x_{1n}) = (z_{11}, z_{12}, \ldots, z_{1n}) = \mathbf{z}_1$, and one dichotomous, $\mathbf{x}_2 = (x_{21}, x_{22}, \ldots, x_{2n}) = (z_{21}, z_{22}, \ldots, z_{2n}) = \mathbf{z}_2$. Their distributions were $x_{1i} \sim Normal(0, 1)$ and $x_{2i} \sim Bernoulli(0.5)$, $i = 1, 2, \ldots, n$. In turn, the vector of coefficients related to the covariates on the longitudinal model were $\beta_1 = -2$ and $\beta_2 = 0.5$, whereas the coefficients associated with covariates in the survival sub-model had opposite signs, $\psi_1 = 2$ and $\psi_2 = -0.5$. The vector of random effects followed a bivariate Normal distribution, $\mathbf{b}_i = (b_{0i}, b_{1i}) \sim Normal_2\left(\begin{bmatrix} 0.2 \\ -1.2 \end{bmatrix}, \begin{bmatrix} 1.3^2 & 0.52 \\ 0.52 & 1^2 \end{bmatrix}\right)$, $i = 1, 2, \ldots, n$. In addition to that, the failure times followed a $Gompertz(\alpha = 1.1, \lambda = 0.1)$ distribution and we set that subjects could be followed-up until 3 unities of time. Then, the generated survival times greater than 3 were considered censorings. The parameter that links both sub-models was set to be $\eta = 0.25$. This whole set-up led to an approximate 50% of observed failure times and a maximum of six measurements per person. More details related to the data generation process and about the Gompertz distribution can be seen in Appendix B, page 110.

The fitted models and their notation are described in Table 5. This configuration was based on the idea to adjust the true model, the proposed model, as well as an alternative framework representing an intermediate model in what concerns their flexibility and robustness.

Basically, we can compare the true model for the longitudinal component against the approximation by the BP. In what concerns the survival sub-model, the comparison takes its place between the true distribution (Gompertz) with the Weibull distribution and the BP approximation. It is noteworthy that, in our example, the hazard function of the Gompertz distribution presents an increasing behavior (see the hazard function in page

Table 5 – Description and notation of the fitted models.

Notation	Longitudinal Sub-Model	Survival Sub-Model
$\mathcal{M}_{Go}^{N}$	Linear Normal Mixed Effects	Gompertz
$\mathcal{M}_{We}^{N}$	Linear Normal Mixed Effects	Weibull
$\mathcal{M}_{BP_{m_S}}^{BP_{m_L}}$	Bernstein Polynomial with degree $m_L - 1$	Bernstein Polynomial with degree m_S

We recall the degree m_S is referred to the cumulative baseline hazard function. Therefore, the baseline hazard function is modeled by a BP with degree $m_S - 1$.

110) and that this characteristic is encompassed by the Weibull distribution. Evidently, the theory of the BP indicates that they are also able to capture such function. So, in a sense, all configurations of Table 5 are expected to present a good performance.

The prior distributions for the parameters were set to be weakly informative. These settings were defined aiming at a good mixing of the MCMC chains regarding the total number of parameters and model complexity. There is a set of parameters that are the same for all fitted models, they are $\boldsymbol{\beta}$, σ_ϵ, η, $\boldsymbol{\psi}$. Evidently, the prior distributions for these parameters were the same: $\boldsymbol{\beta} \sim Normal_2(\mathbf{0}_2, 5^2\mathbb{I}_2)$, $\sigma_\epsilon \sim Gamma(0.1, 0.1)$, $\eta \sim Normal(0, 5^2)$ and $\boldsymbol{\psi} \sim Normal_2(\mathbf{0}_2, 5^2\mathbb{I}_2)$. We remind that the notation $\mathbb{I}_2$ represents a 2×2 identity matrix and $\mathbf{0}_2$ is a vector of length 2 in which all the components are equal to zero. In addition, the prior distributions for the parameters of the random effects under a Normal distribution were $\boldsymbol{\mu}_b \sim Normal_2\left(\mathbf{0}_2, 5^2\mathbb{I}_2\right)$ and $\Sigma_b^{-1} \sim Wishart(4, 1/2\mathbb{I}_2)$. Under a Bernstein Polynomial approach with degree $m_L - 1 = 4$, the prior distributions were $\boldsymbol{\mu}_\xi \sim Normal_{m_L}\left(\mathbf{0}_{m_L}, 5^2\mathbb{I}_{m_L}\right)$ and $\Sigma_\xi \sim Wishart(m_L + 2, (1/m_L)\mathbb{I}_{m_L})$. At last, the shape α and scale λ parameters of the Gompertz and Weibull distributions had, *a priori*, a $Gamma(0.1, 0.1)$ distribution. The vector of coefficients of the BP in the survival sub-model was set with $\log(\gamma_k) \sim Normal(0, 5^2)$, $k = 1, 2, \ldots, m_S$. The degree of the BP for the longitudinal sub-model was $m_L = 5$ and for the survival sub-model was $m_S = 5$ and 10. This simulation study was based in 500 Monte Carlo (MC) replicas. Here we set the burn-in as 3,000, a lag of 1, and 4,000 posterior values were saved for the analyzes.

It is important to mention that the shape parameter of the Gompertz distribution is defined in the real domain, *i. e.*, $\alpha \in \mathbb{R}$ (see details in the Appendix B, page 110). Therefore, by setting the prior distribution for this parameter as $\alpha \sim Gamma(0.1, 0.1)$, we are restricting the possible values for this quantity to the $\mathbb{R}^+$. Then, in this case, we are using a prior information that the baseline hazard function presents an increasing behavior. Thus, we give the true model an advantage.

We also tested to fit all these models with a different prior distribution for the variance-covariance matrices, Σ_b and Σ_ξ. In this alternative, we did not take into account

the dependence between the vector of coefficients / random effects. However, note that if $\Sigma_b = \begin{bmatrix} \sigma_{00}^2 & \sigma_{01} \\ \sigma_{10} & \sigma_{11}^2 \end{bmatrix} = \begin{bmatrix} \sigma_{00}^2 & 0 \\ 0 & \sigma_{11}^2 \end{bmatrix}$. That is, if $\sigma_{01} = \sigma_{10} = 0$, we have that $Var[Y_i(t_{ij})] = \sigma_\epsilon^2 + [\mathbf{f}(t_{ij})]\,\Sigma_b\,[\mathbf{f}(t_{ij})]^\top = \sigma_{00}^2 + \sigma_{11}^2 t_{ij}^2 + \sigma_\epsilon^2$. Thus, there can be an identifiability problem with the parameters σ_{00}^2 and σ_ϵ^2 because $\sigma_{00}^2 + \sigma_\epsilon^2 = (\sigma_{00}^2 + c) + (\sigma_\epsilon^2 - c)$, where c can be any real constant. In the case of Bernstein Polynomials, Σ_ξ is a $m_L \times m_L$ matrix; then, if the element $\sigma_{kl} = 0$ for $k \neq l$ we have that

$$\begin{aligned} Var[Y_i(t_{ij})] &= \sigma_\epsilon^2 + \sum_{k=1}^{m_L}\sum_{l=1}^{m_L} b_{k,m_L-1}\left(\frac{t_{ij}}{T_{max}}\right) b_{l,m_L-1}\left(\frac{t_{ij}}{T_{max}}\right)\sigma_{kl} \\ &= \sigma_\epsilon^2 + \sum_{l=1}^{m_L}\left[b_{k,m_L-1}\left(\frac{t_{ij}}{T_{max}}\right)\right]^2 \sigma_l^2. \end{aligned}$$

As a result, at time $t_{i1} = 0$ the variance of $Y_i(t_{i1})$ will be

$$\begin{aligned} Var[Y_i(t_{i1})] &= \sigma_\epsilon^2 + \sum_{l=1}^{m_L}\left[b_{l,m_L-1}\left(\frac{t_{i1}}{T_{max}}\right)\right]^2 \sigma_l^2 \\ &= \sigma_\epsilon^2 + \sigma_1^2, \end{aligned} \tag{4.4}$$

because the vector of Bernstein basis for the longitudinal sub-model at time $t_{i1} = 0$ is $\mathbf{b}_{m_L-1}(0) = (1, 0, \ldots, 0)$. So, there may have an exchange of information between the variance terms.

The prior distributions for the variance-covariance matrices when we did not take the correlation into account were $\Sigma_b = \text{diag}(\sigma_{b_0}^2, \sigma_{b_1}^2)$, $\sigma_{b_l} \sim Gamma(1, 1)$, $l = 0, 1$ and $\Sigma_\xi = \text{diag}(\sigma_{\xi_1}^2, \sigma_{\xi_2}^2, \ldots, \sigma_{\xi_{m_L}}^2)$, $\sigma_{\xi_l} \sim Gamma(1, 1)$, $l = 1, 2, \ldots, m_L$. Results for this test can be seen in Appendix B (page 111).

A solution to the identifiability matter discussed above is to fully model the variance-covariance matrix. It can be done by assuming that they follow, *a priori*, a Wishart distribution. This specification avoids the exchange of information between the variances since the structure of these matrices are sampled altogether, as a single parameter. Therefore, this is the justification for the usage of this prior distribution.

Table 6 shows the coverage percentage (CP) for the main parameters based on the HPD interval with 95% probability. The dotted line indicates a separation between the results of the parameters associated to the longitudinal sub-model and those related to the survival part. The dashes in this table mean that there is no related parameters in the BP model.

Focusing on the first part of Table 6, note that the probability that the true value of a parameter is in the HPD interval is extremely similar for all four fitted models. Moreover, most of them are around the nominal value of 95%. This result was more than expected for

Table 6 – Coverage percentage based on HPD intervals for main the parameters.

	$\mathcal{M}^N_{Go}$	$\mathcal{M}^N_{We}$	$\mathcal{M}^{BP_5}_{BP_5}$	$\mathcal{M}^{BP_5}_{BP_{10}}$
μ_{b_0}	94.00	94.00	95.40	94.60
μ_{b_1}	96.60	95.60	-	-
β_1	94.60	94.20	94.60	94.80
β_2	96.00	96.40	96.00	96.00
σ_{00}	93.60	94.40	-	-
σ_{11}	93.20	93.40	-	-
σ_{01}	95.00	94.60	-	-
σ_ϵ	95.20	96.00	86.20	86.40
ψ_1	96.60	52.40	96.20	96.40
ψ_2	94.40	93.40	94.80	94.60
η	96.80	61.80	96.60	96.60

the first ($\mathcal{M}^N_{Go}$) and the second ($\mathcal{M}^N_{We}$) models because, in these cases, we are fitting the exact true model. Nonetheless, we can see that approximating the time-varying behavior of the longitudinal variable with Bernstein Polynomials - regardless of the degree of the BP in the survival sub-model - showed to be no disadvantageous for the most important parameters. The only parameter in which the CP is far from the nominal value is the standard deviation of the measurement error σ_ϵ. In this case, the approximation by BP covered the true value only on 86% of the times.

The last three rows of Table 6 show results of the CP for parameters related to the survival sub-model. It is clear that assuming a Weibull distribution for the survival times can present issues. Using this model, the coverage for the coefficient related to the continuous covariate ψ_1 was only of 52.40%; for η, which is the most important parameter in this framework, the model $\mathcal{M}^N_{We}$ covered the true value only in 61.80% of the times. On the other hand, the true model and the approximation by BP, with both degrees evaluated, were able to cover the true values of these parameters in percentages close to the nominal

value.

In order to further investigate the model performances regarding each of the main parameters, Table 7 shows the mean (and the standard deviation) of the relative biases (RBs) for all fitted models. This measure evaluates the difference between the point estimates and the true values of the parameters, taking into account their scale. These quantities were based on the posterior means, medians and modes of the 500 MC replications. The formula for the RB measure can be seen in Equation (4.5).

$$RB(\theta) = \frac{\hat{\theta} - \theta_{true}}{|\theta_{true}|} 100\%, \tag{4.5}$$

where θ represents an unknown parameter. Then, $\hat{\theta}$ is an estimate for this parameter. As mentioned above, the estimates we used here were the posterior mean, median and mode. At last, θ_{true} is the true value for this quantity.

Table 7 – Mean and standard deviations of the relative biases for the main parameters based on the posterior means, medians and modes.

	Mean				Median				Mode			
	$\mathcal{M}_{Go}^{N}$	$\mathcal{M}_{We}^{N}$	$\mathcal{M}_{BP_5}^{BP_5}$	$\mathcal{M}_{BP_{10}}^{BP_5}$	$\mathcal{M}_{Go}^{N}$	$\mathcal{M}_{We}^{N}$	$\mathcal{M}_{BP_5}^{BP_5}$	$\mathcal{M}_{BP_{10}}^{BP_5}$	$\mathcal{M}_{Go}^{N}$	$\mathcal{M}_{We}^{N}$	$\mathcal{M}_{BP_5}^{BP_5}$	$\mathcal{M}_{BP_{10}}^{BP_5}$
μ_{b_0}	2.80 (62.69)	0.78 (62.61)	5.03 (67.41)	4.89 (67.34)	2.79 (62.74)	0.74 (62.66)	5.04 (67.52)	4.95 (67.36)	3.52 (63.60)	0.80 (63.31)	5.28 (68.44)	5.12 (68.32)
μ_{b_1}	0.20 (6.35)	-1.34 (6.41)	- -	- -	0.17 (6.34)	-1.38 (6.41)	- -	- -	0.16 (6.37)	-1.49 (6.45)	- -	- -
β_1	0.18 (4.42)	0.39 (4.45)	0.22 (4.49)	0.22 (4.48)	0.18 (4.42)	0.39 (4.45)	0.22 (4.49)	0.21 (4.48)	0.16 (4.47)	0.36 (4.46)	0.21 (4.50)	0.20 (4.53)
β_2	-0.03 (33.24)	-0.21 (33.25)	-1.06 (33.18)	-0.98 (33.14)	-0.04 (33.24)	-0.26 (33.26)	-1.03 (33.19)	-0.96 (33.15)	-0.14 (33.77)	-0.12 (33.48)	-1.01 (33.77)	-0.65 (33.76)
σ_{00}	-3.08 (12.94)	-2.35 (13.02)	- -	- -	-3.67 (12.92)	-2.95 (13.00)	- -	- -	-4.73 (12.94)	-4.08 (13.07)	- -	- -
σ_{11}	-1.05 (13.35)	-0.84 (13.36)	- -	- -	-1.86 (13.29)	-1.64 (13.31)	- -	- -	-3.37 (13.27)	-3.02 (13.32)	- -	- -
σ_{01}	4.65 (24.61)	3.53 (24.79)	- -	- -	4.95 (24.51)	3.88 (24.70)	- -	- -	5.44 (24.65)	4.61 (24.83)	- -	- -
σ_{ϵ}	0.50 (2.30)	0.33 (2.29)	-2.06 (2.39)	-2.06 (2.39)	0.46 (2.31)	0.29 (2.29)	-2.08 (2.39)	-2.09 (2.39)	0.37 (2.35)	0.23 (2.32)	-2.14 (2.40)	-2.13 (2.40)
ψ_1	1.00 (6.24)	-11.75 (5.91)	1.52 (6.32)	1.27 (6.28)	0.90 (6.23)	-11.82 (5.90)	1.43 (6.31)	1.19 (6.27)	0.71 (6.31)	-11.98 (5.94)	1.20 (6.32)	1.03 (6.30)
ψ_2	1.20 (27.08)	13.83 (24.62)	0.13 (27.17)	-1.16 (27.02)	1.29 (27.08)	13.87 (24.60)	0.19 (27.16)	-1.08 (27.00)	1.60 (27.25)	13.82 (24.79)	0.28 (27.33)	-1.07 (27.10)
η	0.90 (12.54)	-22.43 (12.17)	1.49 (12.61)	1.42 (12.51)	0.74 (12.51)	-22.54 (12.15)	1.34 (12.58)	1.25 (12.46)	0.51 (12.55)	-22.73 (12.25)	1.22 (12.51)	0.94 (12.46)

We can observe from this table that the RBs for the parameters related to the longitudinal sub-model were substantially low (< 6) for all models - regardless of the posterior point estimates. Again, this conclusion is almost immediate for the models $\mathcal{M}_{Go}^{N}$ and $\mathcal{M}_{We}^{N}$ because they are the true ones. The novelty that we point out is the good performance of the BP. Furthermore, it is important to highlight that although the CP for the parameter σ_ϵ was somewhat lower than the nominal value using models $\mathcal{M}_{BP_5}^{BP_5}$ and $\mathcal{M}_{BP_{10}}^{BP_5}$, the RBs are considerably low. It means that the estimates for this parameter are really close to the true value, but the HPD intervals are too short to be capable to cover it.

When it comes to the parameters of the survival sub-model the RBs values were low only for the true model and the approximation by BP, both for $m_S = 5$ and $m_S = 10$. The RBs for these parameters when the survival times were assumed to follow a $Weibull$ distribution were relatively higher when compared to the others. Moreover, we highlight that these elevated values for model $\mathcal{M}_{We}^{N}$ occurred especially for the parameter that links both sub-models η. This is the most important parameter in the joint models.

The values of the CP and the RBs presented so far indicate some interpretations. Nonetheless, a more complete analysis can be done by observing the entire distribution of the RBs. For this purpose, Figures 9 to 15 show the boxplots of the relative biases based on all the 500 posterior means, medians and modes. It is important to mention that we removed the outliers from these figures. The aim with this removal was to shorten the intervals in order to better evaluate the results. These same figures with all values represented can be seen in Appendix B, page 112. In addition to that, one should be attentive to the different scales of these figures.

Figures 9 to 12 show results of the parameters related to the longitudinal sub-model. One more time, we anticipate that it is expected that the performance of the first two models to be better, since they are exactly the model of which the longitudinal data was generated. It is possible to note from Figure 9 that the estimates for the intercept μ_{β_0} presented a quite large variation. This variation was a little bit greater for the BP approximation. Nevertheless, the medians in these boxplots are all close to 0, which indicates non-biased estimates. The boxplots for the approximation via BP presented a slightly higher median for this parameter, but still close to zero.

Figure 10 focuses on the parameter β_1. It represents the coefficient related to the continuous covariate in the longitudinal sub-model. For this case, we can see that all the medians are close to zero with a minor elevation for the models $\mathcal{M}_{Go}^{N}$ and $\mathcal{M}_{We}^{N}$. Therefore, the results indicate unbiased estimates. Also, the scale of this figure is remarkably lower when compared to the previous one.

The next figure concerns the coefficient associated to the binary covariate, β_2. We can observe in Figure 11 that the estimates presented a somewhat large variation, but the median for all the RBs were close to zero. Thus, we have unbiased estimates for this

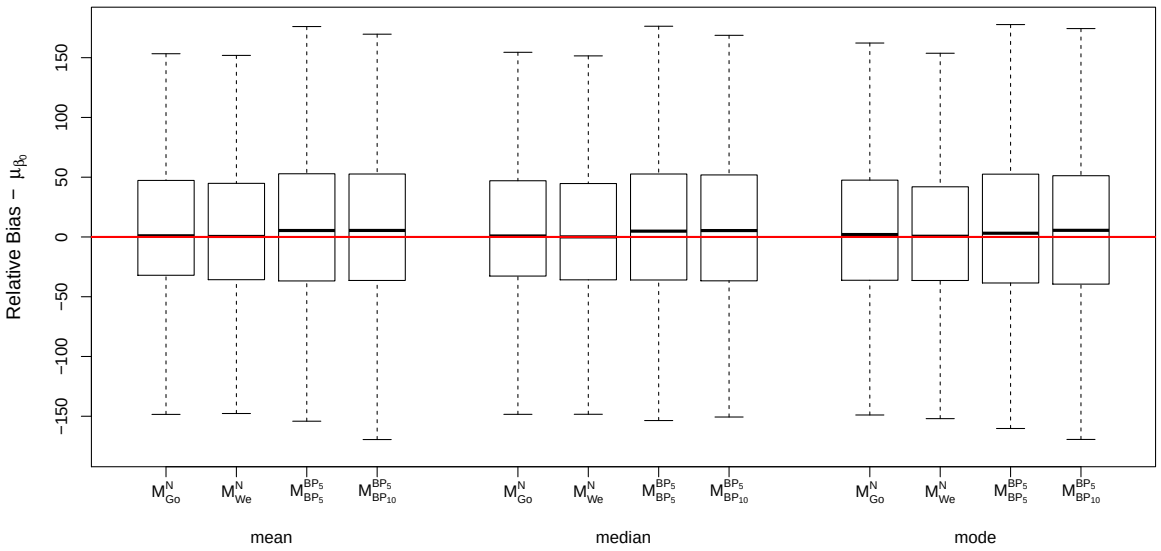


Figure 9 – Comparison of the relative biases of the parameter μ_{b_0} based on the posterior mean, median and mode and comparing each modeling approach.

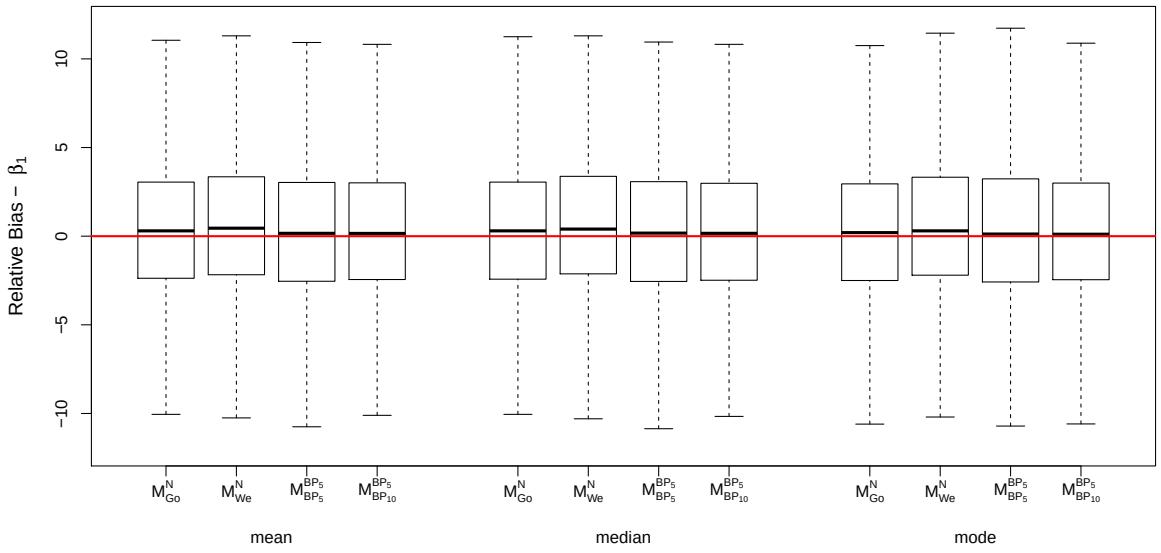


Figure 10 – Comparison of the relative biases of the parameter β_1 based on the posterior mean, median and mode and comparing each modeling approach.

parameter.

The results related to the standard deviation of the measurement error are shown in Figure 12. Note that the scale of this figure is the shorter comparing to the figures in this section. The models $\mathcal{M}^N_{Go}$ and $\mathcal{M}^N_{We}$ present a better performance compared to that of the BP approximation. However, we point out that the distribution of the RBs obtained by using the BP approximation is concentrated in the interval $(-5, 0)$. Thus, despite of this deviation, presented in the BP approximation from what could be considered unbiased, its

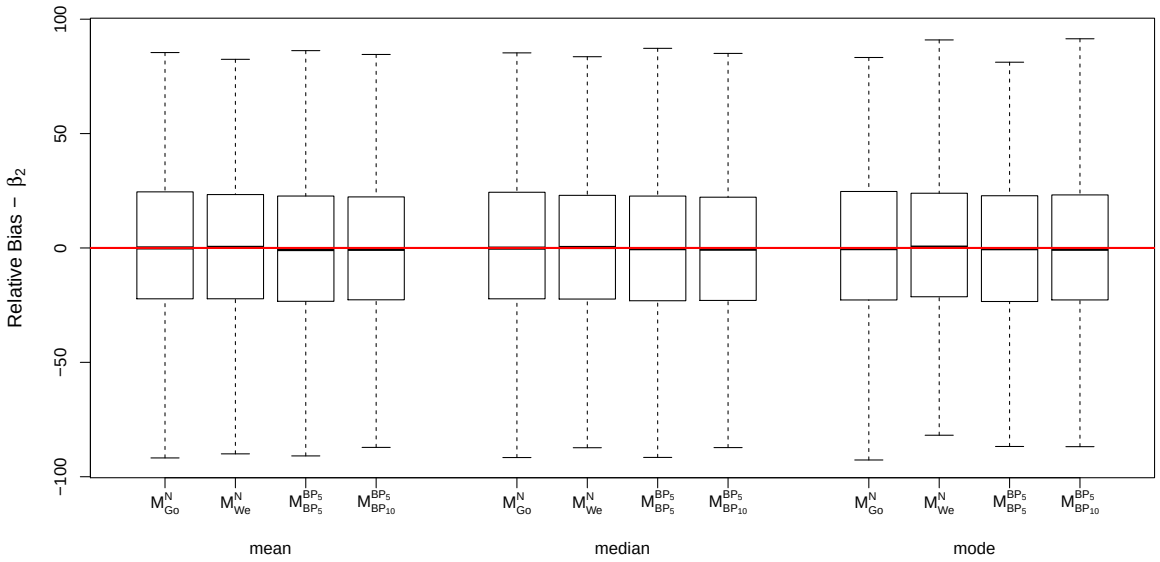


Figure 11 – Comparison of the relative biases of the parameter β_2 based on the posterior mean, median and mode and comparing each modeling approach.

values are still reasonable. As a conclusion, it reinforces the interpretation that we have estimates that are close to the true value but with small HPD intervals.

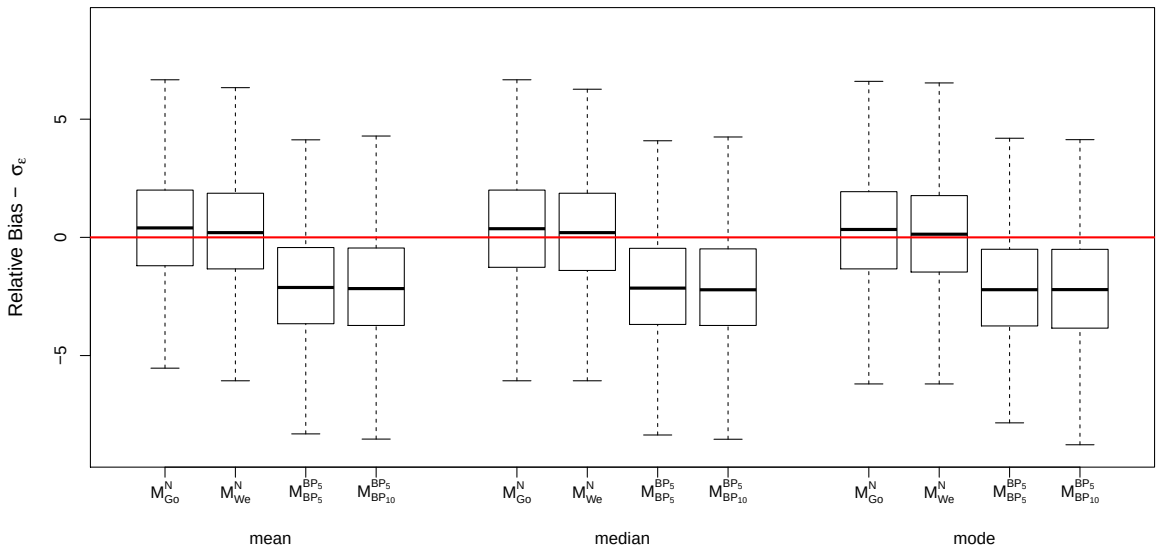


Figure 12 – Comparison of the relative biases of the parameter σ_ϵ based on the posterior mean, median and mode and comparing each modeling approach.

The next figures (13 to 15) show results of parameters related to the survival sub-model. In Figure 13 we can see the relative biases for the parameter ψ_1. This is the representation of the coefficient related to the continuous covariate. We can observe that to assume the survival times with a Weibull distribution, *i. e.* to use the model $\mathcal{M}^N_{We}$, leads to an underestimation of this parameter. On the other hand, the estimates obtained

by modeling the baseline hazard function with Bernstein Polynomials indicates that it is a competitive alternative. We recall that there is an "advantage" in the model $\mathcal{M}^N_{We}$ since the longitudinal sub-model is the same as the generated one.

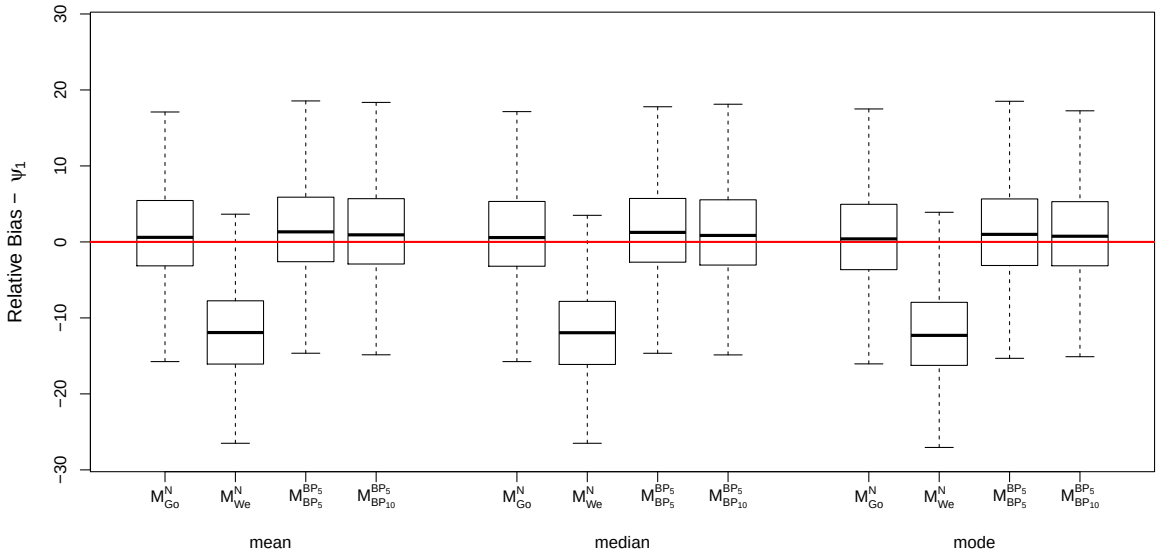


Figure 13 – Comparison of the relative biases of the parameter ψ_1 based on the posterior mean, median and mode and comparing each modeling approach.

Considering the binary covariate on the survival sub-model, see Figure 14. It shows the relative biases for all fitted models and based on three point estimates. Similarly to the case in Figure 13, this parameter is overestimated for the Weibull sub-model. The performance of the BP with both degrees considered and the true model are in accordance. That is, low bias and similar variability.

Finally, the most important parameter in the joint model framework is the linking parameter η. We can see in Figure 15 the performance of this parameter for all fitted models. Note that the estimates for η using the model $\mathcal{M}^N_{We}$ are underestimated. The true model $\mathcal{M}^N_{Go}$, of course, presents good results as well as both models using Bernstein Polynomials.

A form of comparing these four models can be via usual comparison measures. Hence, Figure 16 shows the boxplots with the values of DIC, -2LPML and -2WAIC relative to the values of the true model. That is, for each MC sample we calculated the ratio between a particular comparison measure for one of the models with respect to the true model. These values compose the boxplots in this figure. Then, the interpretation of the relative rescaled comparison measures is, within the fitted models, *the lower the better*. Still considering the relative measures, an additional interpretation is that when these values are lower than 1 it means that, in this case, the performance of this model was better than the true one. We also removed outliers from these boxplots. A figure with the

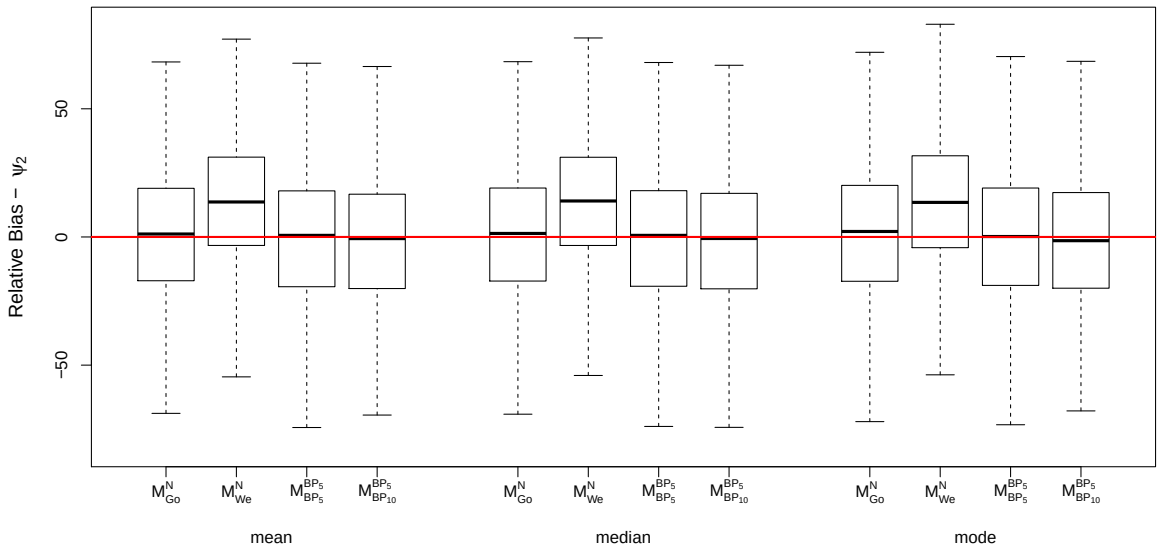


Figure 14 – Comparison of the relative biases of the parameter ψ_2 based on the posterior mean, median and mode and comparing each modeling approach.

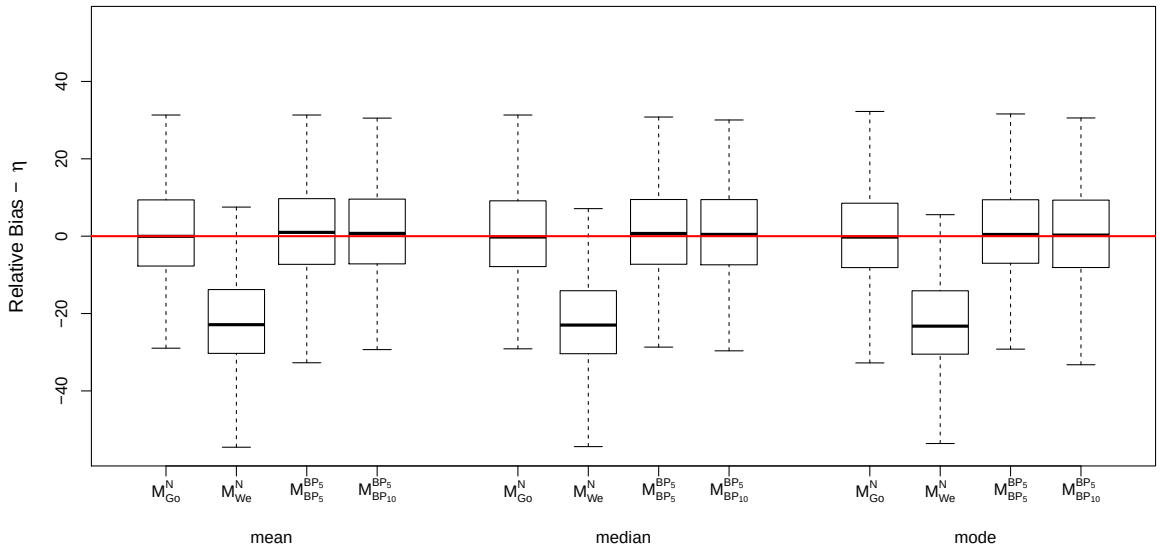


Figure 15 – Comparison of the relative biases of the parameter η based on the posterior mean, median and mode and comparing each modeling approach.

comparison measures for all models and another for Figure 16 with the outliers can be seen on the Appendix B (page 112).

According to the three comparison measures, we can conclude that the best model is the BP approximation with degree $m_S = 5$ for the survival sub-model. Yet, the performance of the model $\mathcal{M}^{BP_5}_{BP_{10}}$ was very similar to the model $\mathcal{M}^{BP_5}_{BP_5}$. However, the comparison measures indicated that the additional flexibility and complexity provided by the BP with degree $m_S = 10$ is superfluous. Furthermore, note that the boxplots of the

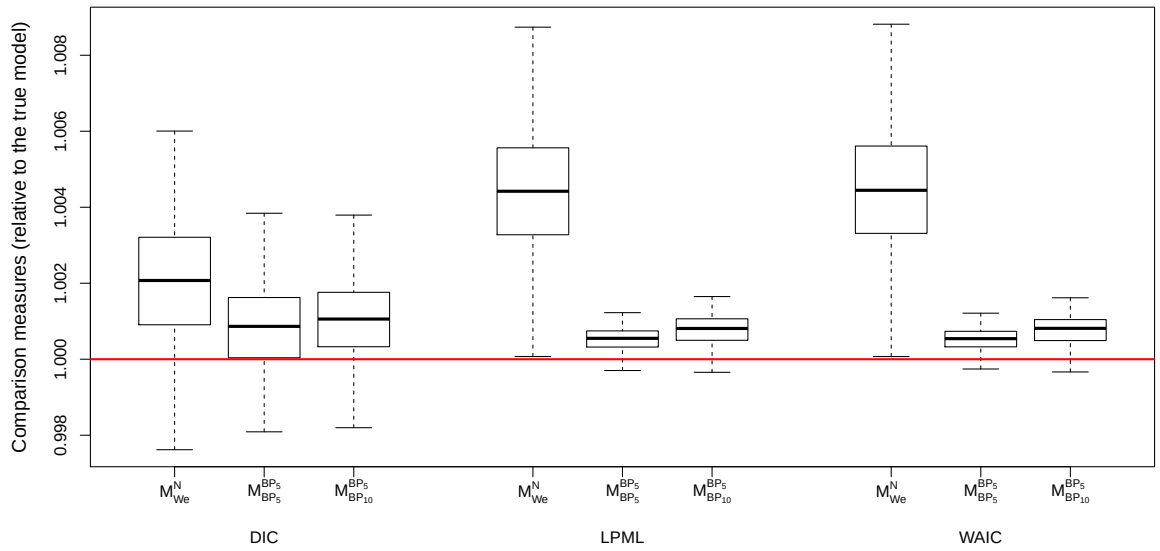


Figure 16 – Comparison measures relative to the true model.

relative LPML and WAIC measures for model $\mathcal{M}^{N}_{We}$ are above the value 1. In Figure 27b we can see that there were few outliers below this value. Hence, this means that the model $\mathcal{M}^{N}_{We}$ had a worse performance compared to the true one $\mathcal{M}^{N}_{Go}$, for almost every single generated data set.

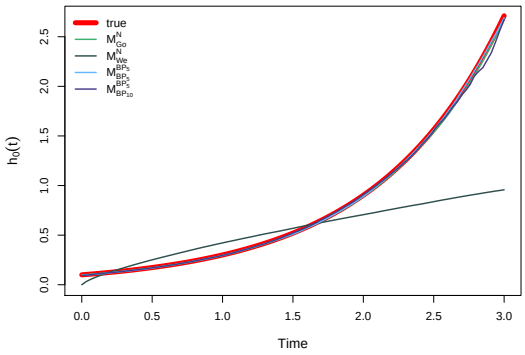
Next, we computed the frequency (and the percentage) of the times that each of the models was chosen as the best one, for every single generate data set. Here, we did not take into account the true model $\mathcal{M}^{N}_{Go}$. This frequency was calculated based on the three comparison measures. These results can be seen on Table 8.

Table 8 – Frequency and percentage of the times in which each model was chosen as the best one - excluding the true model $\mathcal{M}^{N}_{Go}$.

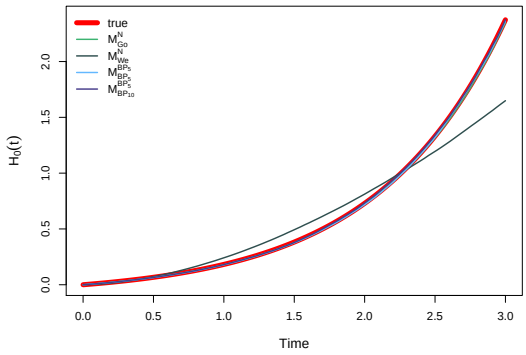
	$\mathcal{M}^{N}_{We}$	$\mathcal{M}^{BP_5}_{BP_5}$	$\mathcal{M}^{BP_5}_{BP_{10}}$
DIC	91 (18.20%)	244 (48.80%)	165 (33.00%)
LPML	2 (0.40%)	418 (83.60%)	80 (16.00%)
WAIC	2 (0.40%)	418 (83.60%)	80 (16.00%)

Then, we can verify that the DIC pointed out to the model $\mathcal{M}^{BP_5}_{BP_5}$ as the best one in almost 50% of all data sets. In the second place, we have the model $\mathcal{M}^{BP_5}_{BP_{10}}$, which was chosen in 33% of the cases. Regarding both LPML and WAIC, this frequency/percentage increases for nearly 84%. As a conclusion, if we had fitted only the models $\mathcal{M}^{N}_{We}$, $\mathcal{M}^{BP_5}_{BP_5}$, and $\mathcal{M}^{BP_5}_{BP_{10}}$, the best choice would be to use the BP in both longitudinal and survival sub-models. In the latter with degree $m_S = 5$. This conclusion is in accordance with that of Figure 16. An additional result is this same frequency (and percentage), but considering all fitted models; see the Appendix, Table 19.

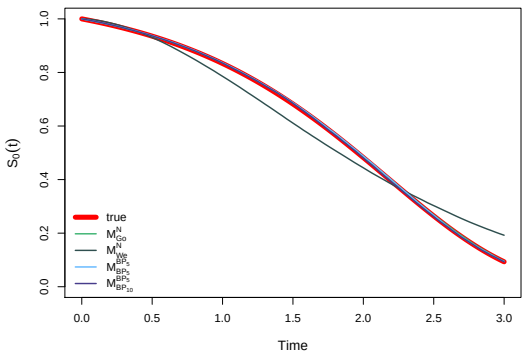
Proceeding with the analysis of the simulated data, Figure 17 shows the mean of the median estimated functions along with their true counterparts. That is, this figure shows the mean of the 500 median estimated functions in the MC scheme. The functions we evaluated were: the baseline hazard function (Panel 17a), cumulative baseline hazard function (Panel 17b), baseline survival function (Panel 17c), and overall mean function (Panel 17d). By observing these panels it is clear that, except for the model $\mathcal{M}^N_{We}$, all fitted models were able to approximate well the survival functions and the overall mean function.



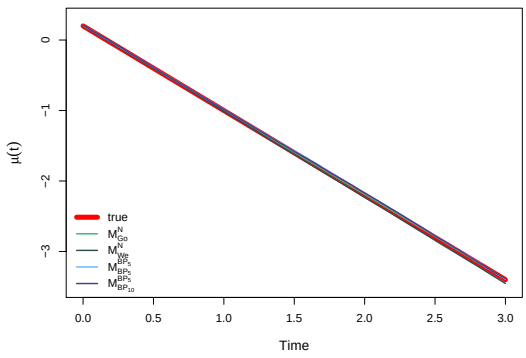
(a) $h_0(t)$.



(b) $H_0(t)$.



(c) $S_0(t)$.



(d) $\mu(t)$.

Figure 17 – Comparison between the median baseline hazard function, cumulative baseline hazard function, baseline survival function and overall mean curve along with the true curve. MC scheme with 500 replications (we summarize the result by taking the mean of the 500 estimated functions).

As a final comment, we expected that the true model would provide the best fit, immediately followed by the model $\mathcal{M}^N_{We}$. This expectation is based on the fact that the longitudinal model is the same as the generated one; and the Weibull distribution encompasses the increasing behavior presented by the true baseline hazard function in this example. However, results show that Bernstein Polynomials were able to perform as good as the true model - except, maybe, for the parameter σ_ϵ. It is also worth pointing out the

importance of accounting *a priori* the information from the correlation of the measurements from the same subjects. That is, to set a prior distribution for the variance-covariance matrices Σ_b and Σ_ξ other than a diagonal matrix. The results shown in the Appendix B (page 111) confirmed that the performance worsened with respect to this simpler prior formulation.

The next section concerns a brief discussion involving the estimated baseline survival function and the Kaplan-Meier estimates.

4.2.1 Difference between estimated baseline survival function and Kaplan-Meier estimates

In this section we bring up a discussion about a result that we found worth of commenting. This result concerns the difference between the estimated baseline survival function $S_0(.)$ obtained via the joint model framework, and the Kaplan-Meier estimates. We elaborated this brief section of debate because of the lack of information about it in papers and books - to the best or our knowledge.

After we had fully implemented the joint model and we obtained the first results, we compared the estimates for the baseline survival function with the KM estimates. At that time, we were expecting that there would be a strong similarity between both results. In our perception back then, this would be an indication that the joint model was correctly implemented. However, it was not what we have encountered. Then, we were caught in a doubt between the belief that: (i) the presence of a possible error in the coding, or (ii) the expected result does not make sense. The primary tentative to solve this puzzle was to search a reinforcement in literature. Although it may be a simple comparison, we were not able to find works with this discussion.

Then, in order to evaluate this difference, we came up with a way of testing this questioning. This test consisted on setting the variance of the prior distribution for the coefficients of the survival sub-model, η and $\boldsymbol{\psi}$, to a value close to zero. The idea of these tests was to use the same `Stan` model and `R` script, changing only the prior variance according to the tests. This was a way of verifying that the difference between the two estimated functions was not due to code error.

We chose randomly one of the 500 generated data sets. The chosen one was the 66^{th} data set. Next, we fitted the true model as well as the best one according to the conclusion of the simulation study. Then, the models were $\mathcal{M}_{Go}^{N}$ and $\mathcal{M}_{BP_5}^{BP_5}$ (see their description in Table 5, page 69). The prior distributions and their respective specifications were all the same, except for the cases described below

1. Test 0: both η and each component of $\boldsymbol{\psi}$ followed a $Normal(0, 5^2)$. This was the

previous configuration;

2. Test 1: the prior distribution for the linking parameter was $\eta \sim Normal(0, 0.0001^2)$;
3. Test 2: the prior distribution for each component in the vector of coefficients $\boldsymbol{\psi}$ was $Normal(0, 0.0001^2)$;
4. Test 3: both η and all coefficients of the vector $\boldsymbol{\psi}$ followed, *a priori*, a $Normal(0, 0.0001^2)$ distribution;

The exact difference between each test and the original specification is highlighted in red. In addition, the MCMC configurations were all the same as in the simulation study.

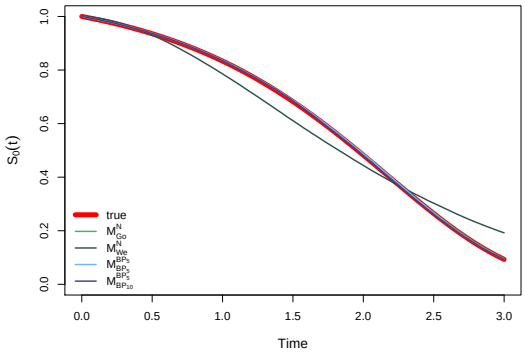
Figure 18 shows the results of the tests described above. In Panel 18a we repeated the result of the simulation study with the regular prior specifications, merely for comparison purposes. This was Test 0. Then, in Panel 18b we can observe results of Test 1. In turn, Panel 18c shows results for Test 2 and, at last, Panel 18d refers to the outcomes of Test 3. In Panels 18b to 18d, the straight black line is the KM estimate and the red curve is the theoretical true baseline survival function. The other curves represent the median estimated curve for $S_0(.)$ and the HPD interval for the fitted models.

The first aspect that we call attention to is the discrepancy between the true curve for $S_0(.)$ (red line) and the estimation by the KM (black line). Then, as we set the information coming from covariates to null, the estimated baseline survival curve has the tendency to get closer to the KM estimate. This feature is more evident in the result of Test 3 (Panel 18d). It makes sense considering that the longitudinal variable is also a function of the time. Therefore, by setting the importance of this information to zero, this other information of time cease to exist.

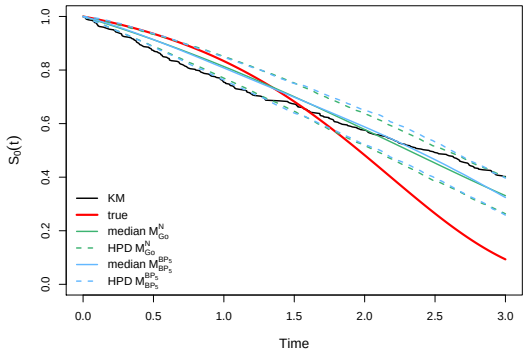
Another point to consider is that the example of our simulation study was not a great choice to clarify the point we came up with. Our guess is that the coefficient of the linking parameter, combined with the values of the longitudinal variable, was too small to be able to show a clear difference. Then, we will do these tests again in the real data set. Nonetheless, we reinforce that by simply comparing the true baseline survival curve and the KM estimate we can see that they do not match.

The conclusion of this experiment is that, under a joint model framework, we should not directly compare the estimated baseline survival function with the estimation obtained using the Kaplan-Meier. Both approximated functions are valid but they deliver different interpretations.

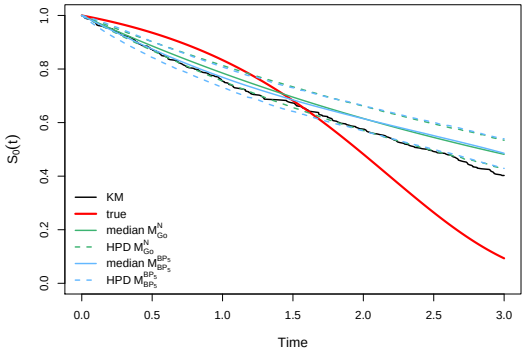
The next chapter shows an application of joint modeling longitudinal and survival data for HIV+ patients.



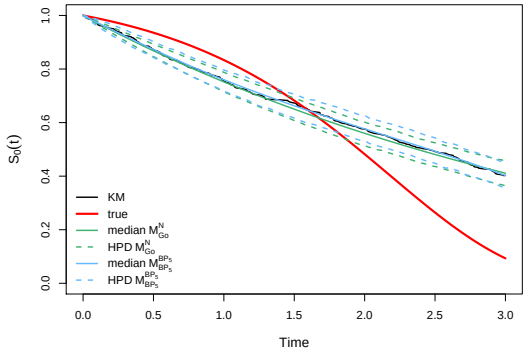
(a) Test 0 - previous result.



(b) Test 1.



(c) Test 2.



(d) Test 3.

Figure 18 – Comparison between the estimation of the baseline survival function $S_0(.)$ obtained in the joint model framework and by the Kaplan-Meier estimator.

5 Real Data Application

This chapter shows an application with real data. The data set we will use in here was collected aiming at comparing two alternative treatments - didanosine (ddI) and zalcitabine (ddC) - for HIV positive patients who failed or were intolerant to zidovudine (AZT). It was first described in Abrams et al. (1994) and it is available in the `R` package `JM` (Rizopoulos, 2010). The response variable was the time, in months, until death. The covariates were: (i) **drug** - the alternative drug used on the treatment, which could be either ddC or ddI; (ii) **gender** - female or male; (iii) **prevOI** - opportunistic diseases at study entry, *i. e.*, if they had AIDS or not at the time they entered in the study; (iv) **AZT** - the reason for their entrance in this study, it could be either because they failed to the treatment with AZT or were intolerant to this drug. In addition to those, a longitudinal variable, the CD4 cell count, was also recorded at five pre-defined points along the follow-up time. They were: the baseline, the 2^{nd}, 6^{th}, 12^{th} and 18^{th} months. Note that, there were at most five measurements of this marker for each patient and the total number of measurements were $N = 1{,}408$. It should also be mentioned that, as seen in Guo and Carlin (2004), the CD4 cell count exhibited skewness; for this reason, this variable will be analyzed in the square root scale.

The analysis of this data set will take place in two parts. In Section 5.1 we describe the data descriptively - this is a fundamental part of any statistical analysis that enhance the understanding of the problem being addressed. Then, in Section 5.2 we fit several models and compare them, reaching to a final result. For all fitted models we run two chains. The burn-in was set to be 5,000, a lag of 1 was taken and, at the end 2,500 posterior values were saved for each chain. We use both the regular comparison measures and the criteria we proposed (page 56) to reach to an optimal degree to choose the best model for this data.

5.1 Descriptive Analysis

This data set is composed of $n = 467$ patients. The minimum and maximum follow-up times were 0.47 and 21.40 months, respectively. The minimum was a failure time, and the maximum was a censorship. The median survival time was 13.20 months and 40.26% of patients died during follow-up period. The remaining ones are right-censored observations.

Table 9 shows descriptive statistics for the baseline covariates. From this table, it is possible to note that 50.75% of patients were treated with the ddC drug. The majority (90.36%) of patients were male, had AIDS diagnosis at study entry (65.74%), and entered

in this study due to AZT intolerance (62.53%).

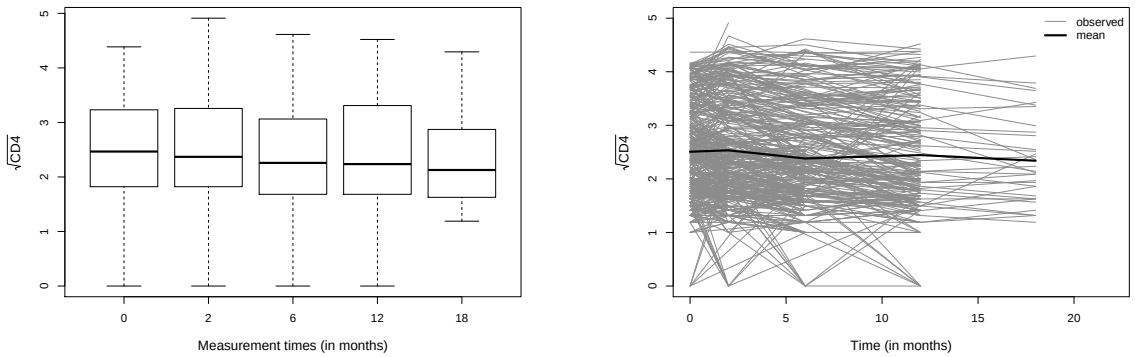
Table 9 – Descriptive statistics of the categorical variables.

Covariate	Frequency (%)
Drug	
zalcitabine (ddC)	237 (50.75%)
didanosine (ddI)	230 (49.25%)
Gender	
female	45 (9.64%)
male	422 (90.36%)
Opportunistic diseases at study entry	
AIDS diagnosis	307 (65.74%)
no AIDS diagnosis	160 (34.26%)
AZT	
failure	175 (37.47%)
intolerance	292 (62.53%)

Regarding the longitudinal variable, Figure 19a shows the boxplots of the square root of the observed CD4 cell count values at each of the five time points. The number of patients composing each boxplot can be checked in Table 10. We can note that there was not much difference in the distribution of this variable among the measurement times. However, the variability of the measurements at the 6^{th} and the 12^{th} months were higher compared to the others. Also, the minimum value for this variable at the 18^{th} month was higher. It makes sense considering that the patients that were still being under follow up until the end of the study can be regarded to as less susceptible to suffer the event of interest. Thus, as they are, in a sense, stronger and healthier, their CD4 cell count is expected to be higher.

Figure 19b presents the profiles graph of the observed square root of CD4 cell count. This type of graph is usual on longitudinal data analysis. It illustrates the observed behavior of the variable for each patient in all measurement times. The black thick line represents the observed mean curve of this variable. Analogously to what we concluded with Figure 19a, it is also possible to note the decay in the number of observations at the last time point (18^{th} month).

An interesting information is the number of observed measurements for each individual and the number of possible measurements. Since CD4 count is an internal variable, see page 28 and/or Rizopoulos (2012), the number of measurements of each patient depends on their respective follow-up time.



(a) Boxplot of square root of observed CD4 cell count at each measurement time point.

(b) Profiles graph showing the behavior of the observed CD4 cell count for each patient.

Figure 19 – Description of the square root of the observed CD4 cell count.

Table 10 – Number of observed and possible observed measurements, at each time point.

Time	Number of observed measurements (%)	Number of possible measurements (%)
Baseline	467 (100.00%)	467 (100.00%)
2^{nd} month	368 (78.80%)	453 (97.00%)
6^{th} month	310 (66.38%)	404 (86.51%)
12^{th} month	226 (48.39%)	318 (68.09%)
18^{th} month	34 (7.28%)	58 (12.42%)

Note that, at the baseline time, CD4 cell count was observed for all 467 patients under study. The second measurement was performed at the second month of follow-up. At this point there were 453 (or 97% of total patients) being accompanied, however CD4 was accounted for only 368 (or 78.80%) of them. This situation, in which the number of observed values was lower than the possible number of measurements remained until the end of follow-up time. At the last measurement time (the 18^{th} month) there were only 58 subjects on the risk group, which represented only 12.42% of the sample - all the other subjects had already been censored or had failed. Nonetheless, CD4 cell count was evaluated for only 34 of these patients. Therefore, the total number of measurements was $N = 1{,}405$ while it could have been 1,700.

At last, Table 11 presents the frequency and the percentage of the number of measurements. Thus, 61 (13.06%) patients had only one value for the CD4 cell count; in these cases, the measurements were necessarily taken at the baseline. Next, 91 (19.49%) had two observed values: the first one at time 0 and the other one could have been taken in any of the other measurement times. 122 (26.12%) individuals had three values of the longitudinal variable. The highest frequency was for the amount of four measurements, 169 (36.19%) patients had their CD4 cell count evaluated four times. The maximum number of measurements was taken for only 24 (5.14%) of the subjects under study.

Table 11 – Frequency and percentage of the number of measurements.

Number of measurements	1	2	3	4	5
Frequency (%)	61 (13.06%)	91 (19.49%)	122 (26.12%)	169 (36.19%)	24 (5.14%)

The next section contains results of the fitted models and the comparisons that enabled us to choose the best one.

5.2 Modeling Approaches

We fitted all models described in Table 5. Likewise the comparison in the simulation study (in Section 4.2, page 66), we aim at contrasting two models for the longitudinal component: a NLME model *versus* BP approximation with degree $m_L - 1$. Regarding the survival sub-model the evaluation is between the Gompertz distribution, Weibull distribution and BP approximation with degree m_S. For the BP approximation, we fixed m_L as 5, which is the maximum number of measurements; in turn, we varied m_S in the set $\{5, 6 \ldots, 15, 22\}$. The last value is the ceiling of the square root of the sample size $\lceil \sqrt{n} \rceil$. This choice was suggested in Osman and Ghosh (2012) as a suitable one for survival data analysis via BP. We remind that the approximation for the cumulative baseline hazard function $H_0(.)$ is via a BP with degree m_S. Nonetheless, the BP approximation for the baseline hazard function $h_0(.)$ is with a degree $m_S - 1$. In addition, we considered $T_{max} = \left\lceil \max_{\substack{i\in\{1,2,\ldots,n\} \\ j\in\{1,2,\ldots J_i\}}} (t_{ij}, u_i) \right\rceil = 22$. The main goal at this point is to verify which model has the best performance.

The equations for the longitudinal sub-model are

$$\begin{aligned} Y_i(t_{ij}) &= W_i(t_{ij}) + \epsilon_i(t_{ij}),\ i = 1, 2, \ldots n\ ,\ j = 1, 2, \ldots, J_i \\ &= \beta_1\ \text{AZT}_i + \beta_2\ \text{Drug}_i + \beta_3\ \text{Gender}_i + \beta_4\ \text{PrevOI}_i + b_{0i} + b_{1i}t_{ij} + \epsilon_i(t_{ij}), \end{aligned}$$

for the NLME model, and

$$\begin{aligned} Y_i(t_{ij}) &= W_i(t_{ij}) + \epsilon_i(t_{ij}),\ i = 1, 2, \ldots n\ ,\ j = 1, 2, \ldots, J_i \\ &= \beta_1\ \text{AZT}_i + \beta_2\ \text{Drug}_i + \beta_3\ \text{Gender}_i + \beta_4\ \text{PrevOI}_i + (\boldsymbol{\xi}_i^{m_L-1})^\top \mathbf{b}_{m_L-1}\left(\frac{t_{ij}}{T_{max}}\right) + \epsilon_i(t_{ij}), \end{aligned}$$

for the BP approximation with degree $m_L - 1$. In both structures, let $\epsilon_i(t_{ij}) \sim Normal(0, \sigma_\epsilon^2)$. The vector of random effects $\mathbf{b}_i$ follows a $Normal_2(\boldsymbol{\mu}_b, \Sigma_b)$ distribution, while the vector of coefficients of the BP had a $\boldsymbol{\xi}_i^{m_L-1} \sim Normal_{m_L}(\boldsymbol{\mu}_\xi, \Sigma_\xi)$ distribution. The prior distributions were the same as those in the second simulation study (Section 4.2, page 66).

So, *a priori*, $\boldsymbol{\beta} \sim Normal(\mathbf{0}_4, 5^2\mathbb{I}_4)$, $\sigma_\epsilon \sim Gamma(0.1, 0.1)$, $\boldsymbol{\mu}_b \sim Normal_2(\mathbf{0}_2, 5^2\mathbb{I}_2)$ and $\boldsymbol{\mu}_\xi \sim Normal_5(\mathbf{0}_5, 5^2\mathbb{I}_5)$. We modeled the inverse of the variance-covariance structure with Wishart distributions. Thus, $\Sigma_b^{-1} \sim Wishart(4, 1/2\mathbb{I}_2)$ and $\Sigma_\xi^{-1} \sim Wishart(7, 1/5\mathbb{I}_5)$.

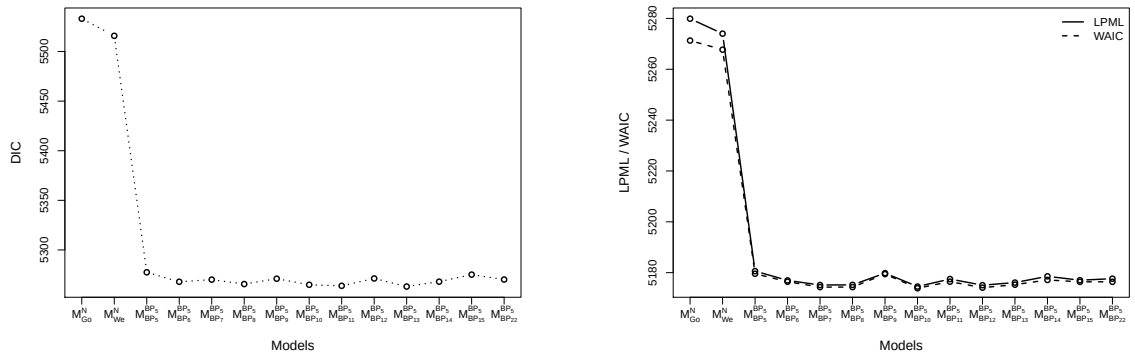
In what concerns the survival sub-model, the equations were

$$\begin{aligned} h(u_i) &= h_0(u_i)\exp\left\{\eta W_i(u_i) + \mathbf{z}_i^\top \boldsymbol{\psi}\right\}, \; i = 1, 2, \ldots, n, \\ &= h_0(u_i)\exp\left\{\eta W_i(u_i) + \psi_1 \text{ AZT}_i + \psi_2 \text{ Drug}_i + \psi_3 \text{ Gender}_i + \psi_4 \text{ PrevOI}_i\right\}, \end{aligned}$$

where the baseline hazard function could be either, (i) $h_0(t) = \lambda e^{\alpha t}$ if we assume that the failure times come from a $Gompertz(\alpha, \lambda)$, (ii) $h_0(t) \approx \left[\sum_{k=1}^{m_S} \gamma_k^{m_S-1} \dfrac{f_{Beta}\left(u/T_{max}; k, m_S - k + 1\right)}{T_{max}}\right]$ if we approximate $h_0(.)$ by BP of order $m_S - 1$, or (iii) or $h_0(t) = \dfrac{\alpha}{\lambda}\left(\dfrac{t}{\lambda}\right)^{\alpha-1}$ if we consider that these times follow a $Weibull(\alpha, \lambda)$ distribution.

For this part, the prior distributions were $\boldsymbol{\psi} \sim Normal_4(\mathbf{0}_4, 5^2\mathbb{I}_4)$ and $\eta \sim Normal(0, 5^2)$. Both parameters of the Weibull distribution α and λ had, *a priori*, a $Gamma(0.1, 0.1)$ distribution. In the case of the Gompertz for the failure times, these distributions were $\alpha \sim Normal(0, 5^2)$ and $\lambda \sim Gamma(0.1, 0.1)$. At last, for the approximation with BP, $\log(\gamma_k^{m_S-1}) \sim Normal(0, 5^2)$, for $k = 1, 2, \ldots, m_S$.

Table 12 shows the comparison measures for each fitted model. It is important to mention that we computed these measures based on the marginal log-likelihood for the longitudinal variable, not the conditional one. The results about the marginal distributions can be verified in page 99. The values of the comparison measures can also be seen in Figure 20. In this figure, Panel 20a shows the values of DIC and Panel 20b displays -2LPML and -2WAIC. We separated the results into two figures due to the scale, facilitating the interpretation and visualization.



(a) DIC. (b) LPML and WAIC.

Figure 20 – Comparison measures for all fitted models.

Table 12 – Comparison measures for fitted models.

Model	DIC	LPML	WAIC	Model	DIC	LPML	WAIC
$\mathcal{M}_{Go}^{N}$	5533.08	5279.89	5271.30	$\mathcal{M}_{BP_{10}}^{BP_5}$	5264.90	**5174.54**	**5173.89**
$\mathcal{M}_{We}^{N}$	5515.83	5274.01	5267.71	$\mathcal{M}_{BP_{11}}^{BP_5}$	5263.82	5177.49	5176.46
$\mathcal{M}_{BP_5}^{BP_5}$	5277.47	5180.61	5179.57	$\mathcal{M}_{BP_{12}}^{BP_5}$	5271.25	5175.08	5174.11
$\mathcal{M}_{BP_6}^{BP_5}$	5267.92	5176.95	5176.44	$\mathcal{M}_{BP_{13}}^{BP_5}$	**5263.11**	5176.10	5175.28
$\mathcal{M}_{BP_7}^{BP_5}$	5270.08	5175.14	5174.33	$\mathcal{M}_{BP_{14}}^{BP_5}$	5268.02	5178.56	5177.13
$\mathcal{M}_{BP_8}^{BP_5}$	5265.62	5175.20	5174.32	$\mathcal{M}_{BP_{15}}^{BP_5}$	5275.15	5177.04	5176.39
$\mathcal{M}_{BP_9}^{BP_5}$	5271.01	5179.82	5179.44	$\mathcal{M}_{BP_{22}}^{BP_5}$	5270.17	5177.63	5176.40

Thus, according to the DIC measure, the best model is the BP approximation with $m_L = 5$ for the longitudinal sub-model and $m_S = 13$ in the survival part. On the other hand, both LPML and WAIC indicated that the best model is the BP approximation with degree $m_S = 10$ for the survival component. Moreover, note that even in the worst scenario of the BP approximation - which happened when $m_S = 5$ for all three measures -, this approach had a better performance when compared to the others. It is simple to check this statement by observing Figure 20. In the panels of this figure, the first and the second results are related to the Gompertz and the Weibull distributions. All the other points are of the Bernstein Polynomials approximation with increasing degrees. We point out to the big drop from the first two points of these figures in relation to the others. Another comment about the choice of the BP degree in the survival sub-model is that the indicated value of $\lceil\sqrt{n}\rceil = \lceil\sqrt{467}\rceil = 22$ seems to be unnecessarily large, as none of these measures select model $\mathcal{M}_{BP_{22}}^{BP_5}$ as the best one.

Another point to consider is that the comparison measures DIC, LPML and WAIC are quite general. In addition to that, the conclusion of the best model that they point out is not in accordance with one another. That is, while DIC indicates $m_S = 13$, both LPML and WAIC lead to the conclusion that it is best to choose $m_S = 10$. Hence, we will use the stopping rule we proposed in this thesis (Section 3.5, page 56) to chose the final model - between all the BP approximations - for the survival component. This stopping rule is specific for the BP approximation.

The results considering the two criteria we propose in this thesis can be seen in

Tables 13 and 14. So, according to the Criterion 1 and both tests used - Sign and Wilcoxon -, the best degree to model the survival component is $m_S = 9$. However, Criterion 2 leads to a doubt between $m_S = 9$ and $m_S = 14$.

Table 13 – Results to the stopping rule for the degree for the BP in the survival sub-model. Criterion 1: difference between coefficients.

	Median		P-value	
Degrees	$\mathbf{D}_{m_S-1}$	$\mathbf{D}_{m_S}$	Sign test	Wilcoxon test
6×7	0.6634	0.4936	< 0.0001	< 0.0001
7×8	0.4936	0.4717	< 0.0001	< 0.0001
$8 \times \mathbf{9}$	0.4717	0.4725	**0.9825**	**0.8345**

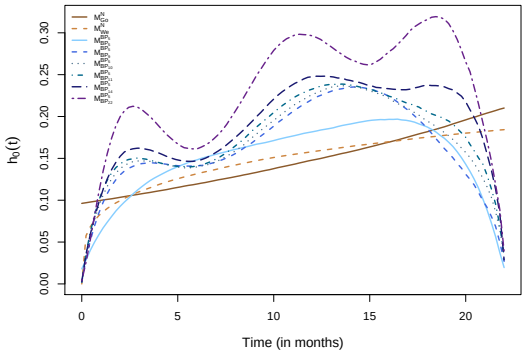
Table 14 – Results to the stopping rule for the degree for the BP in the survival sub-model. Criterion 2: difference between curves

	Median		P-value	
Degrees	$\mathbf{D}_{m_S-1}$	$\mathbf{D}_{m_S}$	Sign test	Wilcoxon test
6×7	0.1522	0.0957	< 0.0001	< 0.0001
7×8	0.0957	0.0793	0.0001	< 0.0001
$8 \times \mathbf{9}$	0.0793	0.0725	**0.1093**	0.0008
9×10	0.0725	0.0658	0.8354	0.0013
10×11	0.0658	0.0607	0.7558	0.0042
11×12	0.0658	0.0536	0.1093	< 0.0001
12×13	0.0536	0.0485	0.4382	0.0002
$13 \times \mathbf{14}$	0.0485	0.0465	0.9538	**0.1550**

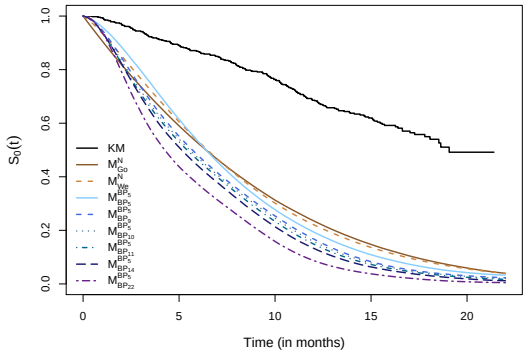
For comparison purposes we plotted the overall mean curve, the baseline hazard function and the baseline survival functions for some of the fitted models. The chosen models were the ones highlighted in red in Table 12, the best models according to our proposed criteria ($m_S = 9$ and $m_S = 14$) as well as the simplest $\mathcal{M}^{BP_5}_{BP_5}$ and the most complex $\mathcal{M}^{BP_5}_{BP_{22}}$ BP approximation. In addition to those, we also plotted these estimated functions for models $\mathcal{M}^{N}_{Go}$ and $\mathcal{M}^{N}_{We}$, as we wanted to show the contrast between BP approximation and other distributions. These graphs are in Figure 21. We emphasize that our intention with this figure is solely to make possible an overall comparison between the approximations based on different models.

By observing Panel 21a we can see the estimations for $h_0(.)$. We emphasize the flexibility of the BP approximation comparing to the results obtained by assuming that the failure times follow a Gompertz or a Weibull distribution. Panel 21b illustrates the baseline survival function for the selected models in comparison to the Kaplan-Meier

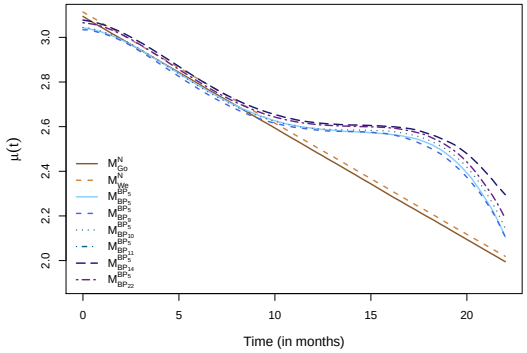
estimator. The curves in this figure reinforces the discussion of Section 4.2 (page 66), that these two results are not directly comparable. Finally, Panel 21c presents the estimates for the overall mean function. We can clearly see that a straight line is not enough to model the variation along the time for this variable. Evidently, we could have used a non-linear structure of time in the longitudinal sub-model. However, the advantage of using the BP to approximate this function is that, even with a small degree, this method was able to approximate the mean curve with a non-linear behavior. And we did not have to assume that this would the behavior of this function.



(a) $h_0(t)$.



(b) $S_0(t)$.



(c) $\mu(t)$.

Figure 21 – Comparison between estimated baseline hazard function, baseline survival function and overall mean curve.

An extra and final comparison between these models is the estimated mean curves based on the selected models of Figure 21c, but along with the trajectory of each subject and the observed mean curve, as in Figure 19b. This comparison can be seen in Figure 28, in Appendix C (page 116). It gives the perspective of the observed values for the longitudinal variable and all the estimated mean curves based on all selected models in the same scale.

We follow our criteria and chose the best model being the BP approximation with

$m_S = 9$ for the survival component. This specification was pointed out in 3 out of the 4 indications in Tables 13 and 14. Point and interval estimates for the vector coefficients related to the covariates in both sub-models ($\boldsymbol{\beta}$ and $\boldsymbol{\psi}$) can be seen in Table 15. For each covariate, if the respective HPD interval includes zero, then the conclusion is that the referred covariate is not statistically related to the response. Otherwise, if this value is not in the interval, we can affirm that there is a relationship between this covariate and the response variable.

By observing Table 15, we can verify that opportunistic diseases at study entry (PrevOI) was important to predict the square root of the CD4 cell count. This variable was lower for those patients that had AIDS at the time they entered the study. More specifically, the mean of the $\sqrt{\text{CD4}}$ count for the patients that had AIDS was -0.8777 unities lower, comparing to those who did not have AIDS. All the other covariates - AZT, drug and gender - do not interfere in this quantity (their HPD intervals had the value 0 included). It is important to mention that according to the medical literature, a low CD4 cell count indicates disease progression. Therefore, the patients that had AIDS presented a worse health status.

Table 15 – Point and interval estimates for the coefficients associated to the covariates of both longitudinal and survival sub-models. Results according to the model $\mathcal{M}_{BP_9}^{BP_5}$.

Longitudinal Sub-model					
Covariate	Mean	Median	Mode	Std. Dev.	HPD 95%
AZT (failure)	-0.0680	-0.0685	-0.0835	0.0944	[-0.2526, 0.1212]
Drug (ddI)	0.0643	0.0645	0.0781	0.0763	[-0.0855, 0.2118]
Gender (male)	0.0402	0.0457	0.0458	0.1458	[-0.2481, 0.3072]
PrevOI (AIDS)	-0.8777	-0.8779	-0.8782	0.0937	[-1.0661, -0.6983]

Survival Sub-model					
Covariate	Mean	Median	Mode	Std. Dev.	HPD 95%
$\sqrt{\text{CD4}}$	-0.9979	-0.9956	-0.9863	0.1232	[-1.2319, -0.7501]
AZT (failure)	0.1263	0.1250	0.1437	0.1666	[-0.1919, 0.4475]
Drug (ddI)	0.3217	0.3192	0.3024	0.1504	[0.0298, 0.6065]
Gender (male)	-0.4050	-0.4099	-0.4054	0.2510	[-0.8890, 0.0996]
PrevOI (AIDS)	0.6385	0.6387	0.6365	0.2349	[0.1859, 1.1128]

Turning our attention to the results of the survival sub-model, we can see that the $\sqrt{\text{CD4}}$ cell count, the alternative drug used in the treatment, and the presence or not of opportunistic diseases at study entry affect the survival to death. On the other hand, the reason for their entrance in this study (AZT - failure or intolerance) and their gender did

not influence this event to happen.

The hazard of death for patients who received ddI as an alternative medication was $e^{0.3217} = 1.3795$ times the hazard of those patients that were prescribed with ddC. In other words, this hazard was 37.95% higher for subjects who received ddI.

In the same way as in the longitudinal sub-model, the patients who had opportunistic diseases at study entrance presented a worse condition. The hazard of death for those patients who had AIDS was $e^{0.6385} = 1.8936$ times the hazard of experimenting this event for the patients who did not present opportunistic infection at the time they entered in the study. It means a 89.36% higher hazard.

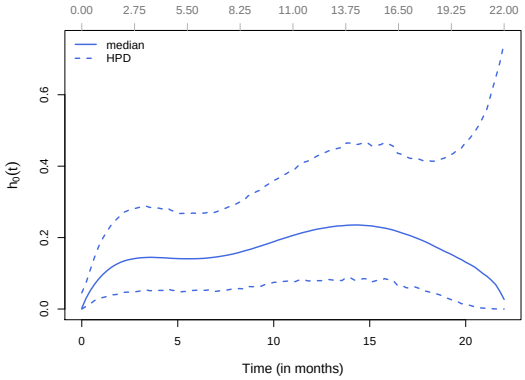
At last, the longitudinal variable $\sqrt{\text{CD4}}$ was indeed an important prognostic factor for disease progression. For a unity (in the square root scale) *decrease* in this variable, the hazard of death increases in $e^{0.9979} = 2.7126$, *i. e.*, it gets 2.71 times higher.

So, the overall conclusion is that the worst prognosis is when patients were treated with ddI, had AIDS at study entry and low $\sqrt{\text{CD4}}$ count. An additional information concerns the functions approximated by the chosen model, that is $\mathcal{M}_{BP_9}^{BP_5}$, and their HPD intervals. So, in Figure 22, Panel 22a contains the baseline hazard function; Panel 22b the cumulative baseline hazard function. In Panel 22c we can see the baseline survival function along with the Kaplan-Meier estimate. Finally, the overall median of the posterior mean curve of the longitudinal variable can be observed in Panel 22d.

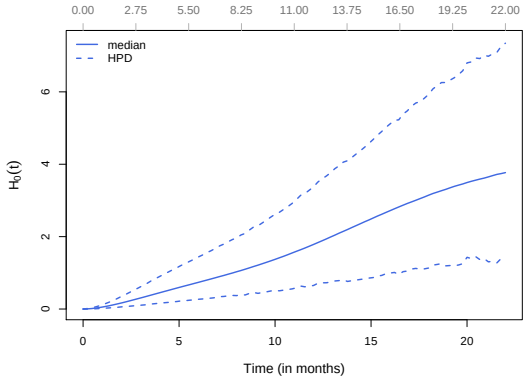
Extra interpretation that can be made by using BP to approximate target functions is the posterior probability that it will change its behavior in a set of time points. This interpretation depends on the posterior sample of the vector of coefficients $\boldsymbol{\mu}_{\xi}^{m_L-1}$ and $\boldsymbol{\gamma}_{m_S-1}$. In practical aspects the calculation is $\mathbb{P}\left((\mu_{\xi_l}^{m_L-1} - \mu_{\xi_{l-1}}^{m_L-1})(\mu_{\xi_{l+1}}^{m_L-1} - \mu_{\xi_l}^{m_L-1}) < 0|\text{Data}\right)$, $l = 2, 3, \ldots, m_L - 2$, for the overall mean function. In the case of the baseline hazard function, this probability is obtained via $\mathbb{P}\left((\gamma_k^{m_S-1} - \gamma_{k-1}^{m_S-1})(\gamma_{k+1}^{m_S-1} - \gamma_k^{m_S-1}) < 0|\text{Data}\right)$, for $l = 2, 3, \ldots, m_S - 2$. We explained a possible way of calculating this probability on page 62. Then, in Table 16 we can observe the results of the estimates of both functions at specific time points as well as the posterior probabilities (p) of a turning point at these times.

Following the results of Table 16, we can see that it is extremely likely that there will be a change in the behavior of the overall mean function at times $t = 11$ months and $t = 16.5$ months. Although it is difficult to visualize this feature at $t = 11$ in Figure 22d due to the scale, we can observe it more clearly in Figure 21c. The change at time $t = 16.5$ months is also easier to observe in the latter mentioned figure. Moreover, we can verify that there was an actual change by examining the estimates for this function in this same table.

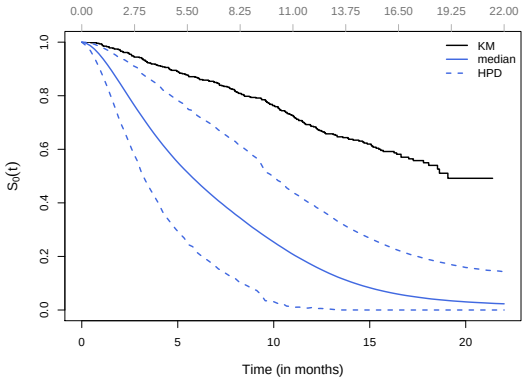
The information of this whole analysis about the behavior of this function has a



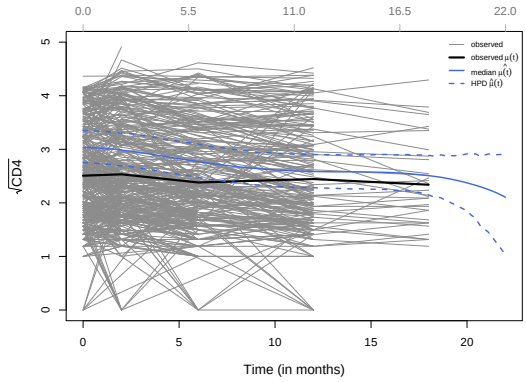
(a) $h_0(t)$.



(b) $H_0(t)$.



(c) $S_0(t)$.



(d) $f(t)$.

Figure 22 – Comparison between estimated baseline hazard function, baseline survival function and overall mean curve. Results according to the model $\mathcal{M}^{BP_5}_{BP_9}$.

Table 16 – Posterior probability of turning points in the overall mean curve and in the baseline hazard function. Results according to the model $\mathcal{M}^{BP_5}_{BP_9}$.

Overall mean curve					
Time	0.00	5.50	11.00	16.50	22.00
$\hat{f}(t)$	3.0342	3.0092	1.9486	3.1703	2.1060
p	-	0.4240	0.9624	0.9518	-

Baseline hazard function									
Time	0.00	2.75	5.50	8.25	11.00	13.75	16.50	19.25	22.00
$\hat{h}_0(t)$	0.0055	0.7751	0.0228	0.0336	0.3981	0.8696	0.1431	0.1344	0.0659
p	-	0.9644	0.5708	0.6056	0.6670	0.8242	0.6852	0.6694	-

practical interpretation. In terms of overall mean and in the square root scale, the CD4 cell count has a decreasing course. This trajectory is maintained until a time around $t = 11$

months. After that, it starts getting higher - possibly due to the efficacy of the treatment. Then, around the 16^{th} month, these counts begin to decrease again. Possible explanations could be a loss of treatment strength or a weakened health, for example.

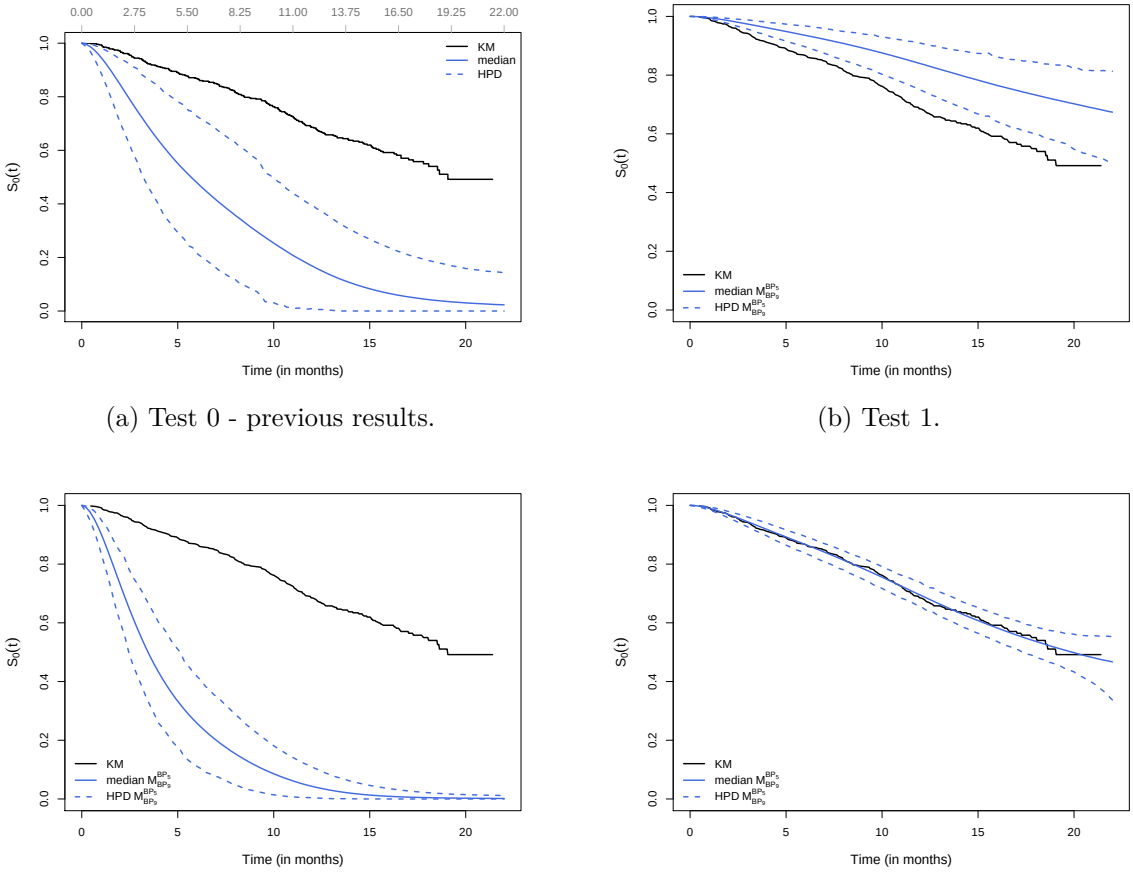
In what concerns the baseline hazard function, results in Table 16 point out that it is very likely that a change will occur in times $t = 2.75$ months and $t = 13.75$ months. By observing Figure 22a, we can see that these points are exactly the ones in which the approximated $h_0(t)$ presents peaks. In addition, since these probabilities for the in between times are around 0.6, we can expect small variations in this curve in the mentioned interval.

As it is an important point, we bring up a brief discussion about the degree. It is true that with a higher m_S the probabilities in the second part of Table 16 would be more accurate, in relation to the time points in which the change will occur. So, for the sake of the accuracy of this specific information, we could verify these probabilities based on a BP model with a higher degree m_S. However, aiming at a good approximation, our criteria pointed out to this value of m_S. Therefore, it is considered adequate and precise enough.

At last, we did again the tests that we have discussed on Section 4.2.1 (page 80) on the chosen model $\mathcal{M}_{BP_9}^{BP_5}$. In summary, the idea was to set the importance of the coefficients in the survival sub-model to null. This procedure was done in three parts. In each part, our goal was to verify the impact of the coefficients in the estimation of the baseline survival function, and to compare it to the KM estimates. The results are shown in Figure 23. In each panel of this figure, the black line represents the KM estimate, the blue thick line is the median baseline survival function, and the dotted line is the HPD interval for this curve. Here, we have used the same prior distributions for the parameters, except, of course, for the case specified in each test. The MCMC specifications were all the same as well.

In Panel 23a of Figure 23, we repeated previous results to easier comparisons. Next, in Panel 23b we can understand what happens with the estimation of the baseline survival function $S_0(.)$ when we ignore the importance of the longitudinal variable. Note that the behavior of this function changes completely, as: (i) it does not go to zero as time increases, and (ii) it gets close to the KM estimates. The results of Test 2 can be seen in Panel 23c. In this case, we can observe that there was not much difference between this approximation and that of Panel 23a. That is, the role of the covariates fixed in time did not affect much the estimation of $S_0(.)$. Finally, when we assumed *a priori* that there was no relationship between the covariates - fixed and varying with time - and the survival response, the final estimation matches exactly the KM estimate. This last result makes sense because, if we have no other information than the survival times (and the indicator of censoring), then the baseline survival function $S_0(.)$ gets equivalent to $S(.)$.

As a conclusion to this discussion, it is very important to take the longitudinal variable into account. In addition, the KM and the baseline survival function in the joint



(a) Test 0 - previous results.

(b) Test 1.

(c) Test 2.

(d) Test 3.

Figure 23 – Comparison between the estimation of the baseline survival function $S_0(.)$ obtained in the joint model framework and by the Kaplan-Meier estimator. Results according to the model $\mathcal{M}^{BP_5}_{BP_9}$.

model scheme have different meanings. However, we should further investigate this matter to fully understand it. The next chapter concerns the conclusions and the ideas for future works.

6 Conclusions and Next Steps

Our main proposition in this thesis was to use Bernstein Polynomials to approximate functions that compose the joint modeling of longitudinal and survival data. We used the BP with degree $m_L - 1$ to model the evolution with time of the longitudinal component, and a BP with degree m_S to approximate the cumulative baseline hazard function. As a consequence, the baseline hazard function is approximated with a BP with degree $m_S - 1$. The usage of the BP to approximate these functions here is one of the contributions of this work.

Throughout this study, we discoursed about the importance of fully considering the information coming from longitudinal and survival data by jointly modeling these components. Regarding the Bernstein Polynomials, we discussed about a diverse list of topics such as: description, properties, historical aspects, explanation of how to use this tool in each of the sub-models in the joint modeling approach, and an intuition of how the BP incorporates data in the approximation.

We implemented the proposed approach in the platform `Stan`. Then, via simulation study, we compared the new method to commonly used approaches. With this study and by real data application, we were able to verify the flexibility of this tool. It can approximate well not only simple functions - such as a linear trends - but also complex ones - like the baseline hazard function in the real application. Besides that, another advantage of using the BP is that there is no need to anticipate the shape of the target function. For example, in the longitudinal sub-model we usually define a form to represent the relationship of the longitudinal variable and the time. This form can be a square root, a simple linear trend, a quadratic or cubic term, among others. When using the BP, it will approximate the true curve in any of the mentioned cases, without having to previously assume one of them. As a result, we can affirm that the proposed modeling approach is in fact a strong competitor.

Another advantage we obtain by using the BP is the posterior probability of a change in the approximated curve. This information is available for $m - 2$ equally spaced points of the domain, where m represent the degree of this polynomial. Such information can be very useful to the researcher, as we can tell when a drug is likely to actually start to increase or decrease a marker, for example.

An additional contribution of this thesis was our proposed method for the degree selection. As we have discussed in previous chapters, this is an important step when using the BP. A degree that is too small leads to an approximated function that can be too smooth. In this case, the resulting approximation may fail to represent important aspects of the target function. On the other hand, if this number is too large, we will have an

unnecessary large number of parameters. This excess can also lead to computational issues. Several works have fixed this quantity based on sample sizes. In opposition to that, our proposal consists in a probabilistic based method that indicates a minimum adequate value. We also presented two stopping rules that indicates the optimal degree. According to the results obtained in the simulation study, our methods have a satisfactory performance. Besides that, our proposal is a robust method, since its usage only needs an indication of a turning point in the function being approximated.

Evidently there are another points we can consider to extend the present work. We will list and discuss some of them.

Initially, we mention that the coding part of this thesis was constructed with the idea of building a package in the software `R`. This tool will certainly facilitate and propagate the usage of our proposals.

We can properly handle missing data. As we have discussed, longitudinal data usually present missing information. Ignoring their presence may lead to bias in the analyses. Therefore, we can focus our attention on treating them in an adequate manner.

Perform residual analysis. We have shown, in the case of the simulated data, that to joint model longitudinal and survival data via BP presents an adequate performance. However, in the case of real data, a residual analysis is indicated to check model adequacy. According to Zhu et al. (2012), non-robust priors for the parameters of the joint model, outliers and model misspecification can lead to biased estimates. In addition, it is known that the occurrence of the event of interest restrain the measurements being taken, *i. e.*, there may exist no more longitudinal observations for a subject after the event or censorship is observed. This phenomena leads to a change in the residual interpretation, since the event occurrence may induce a scenario where there may have more measurements at the beginning of follow-up (Rizopoulos et al., 2010). There are methods for longitudinal and survival data alone, however these methods do not consider the dependence between the components and its implications. Moreover, Rizopoulos et al. (2010) describe examples showing that we can obtain misleading interpretation by using regular residual analysis in joint models. Therefore, this type of modeling requires specific techniques. As a start, we can follow the works of Rizopoulos et al. (2010), Rizopoulos (2012) and Zhu et al. (2012).

We can study different linking forms between the longitudinal and the survival sub-models. Other possibilities of this form may be the acceleration (derivative) of the longitudinal variable, for example. We can also consider more than one longitudinal variable, as well as non continuous longitudinal variables.

Other examples of possible extensions are the consideration of left censoring, interval censoring, informative censoring, presence of a cure fraction and competing risks. Combinations of these proposals are also possible and they can bring interesting results.

Moreover, we can model the survival data in a different structure, such as the proportional odds (Bennett, 1983; Kirmani and Gupta, 2001) and/or the accelerated failure time models.

Another point we can consider to improve the present work concerns the quadrature method that we used to approximate the integral in the survival sub-model. Crowther et al. (2012) says that it is a flaw in this framework not to check if the quadrature form is performing well. Therefore, we can study the usage of different quadrature forms and/or different number of quadratures.

At last, regarding the BP we can plan to use a dynamic prior distribution for the vector of BP coefficients. We have already discussed that there is a straight relationship between the vector of coefficients and the function being approximated. Then, it is possible that this type of prior improves the estimation procedure. We can also assume that the degree of the BP is a random quantity to be estimated. A possible prior distribution could be a Hypergeometric. Then, we could compare these results to our proposed criteria to reach to an optimal degree.

A Details of calculations

In this part of the Appendix we will show the details of the main calculations presented in the text. The results are separated in sections. The order of these sections are in accordance with the organization of this thesis.

Mixed Effects Model

The conditional distribution of the random variable representing the observed values for the longitudinal variable, given the vector of random effects, is Normal, *i. e.* $Y_i(t_{ij})|\mathbf{b}_i \sim Normal(\boldsymbol{\mu}_b, \Sigma_b)$. In addition, we consider that the vector of random effects $\mathbf{b}_i$ also follow a Normal distribution with vector of means $\boldsymbol{\mu}_b$ and variance-covariance matrix Σ_b. Then, the mean and variance of the conditioned distribution of the observed longitudinal variable are given below. The description of this model is in page 22 and 31.

$$\begin{aligned}
\mathbb{E}[Y_i(t_{ij})] &= \mathbb{E}\left[\mathbb{E}[Y_i(t_{ij})|\mathbf{b}_i]\right] = \mathbb{E}[\mathbb{E}[W_i(t_{ij}) + \epsilon_i(t_{ij})|\mathbf{b}_i]] \\
&= \mathbb{E}[\mathbb{E}[W_i(t_{ij}))|\mathbf{b}_i]] + \mathbb{E}[\mathbb{E}[\epsilon_i(t_{ij})|\mathbf{b}_i]] = \mathbb{E}[\mathbb{E}[W_i(t_{ij}))|\mathbf{b}_i]] + \mathbb{E}[\epsilon_i(t_{ij})] \\
&= \mathbb{E}[\mathbb{E}[W_i(t_{ij}))|\mathbf{b}_i]] = \mathbb{E}\left[\mathbb{E}\left[\mathbf{x}_i\boldsymbol{\beta} + \mathbf{f}(t_{ij})\mathbf{b}_i|\mathbf{b}_i\right]\right] = \mathbb{E}\left[\mathbf{x}_i\boldsymbol{\beta} + \mathbf{f}(t_{ij})\mathbf{b}_i\right] \\
&= \mathbb{E}\left[\mathbf{x}_i\boldsymbol{\beta}\right] + \mathbb{E}\left[\mathbf{f}(t_{ij})\mathbf{b}_i\right] = \mathbf{x}_i\boldsymbol{\beta} + \mathbf{f}(t_{ij})\mathbb{E}\left[\mathbf{b}_i\right] \\
&= \mathbf{x}_i\boldsymbol{\beta} + \mathbf{f}(t_{ij})\boldsymbol{\mu}_b
\end{aligned}$$

$$\begin{aligned}
\mathbb{V}ar[Y_i(t_{ij})] &= \mathbb{E}\left[\mathbb{V}ar\left[Y_i(t_{ij})|\mathbf{b}_i\right]\right] + \mathbb{V}ar\left[\mathbb{E}\left[Y_i(t_{ij})|\mathbf{b}_i\right]\right] \\
&= \mathbb{E}\left[\mathbb{V}ar\left[W_i(t_{ij}) + \epsilon_i(t_{ij})|\mathbf{b}_i\right]\right] + \mathbb{V}ar\left[\mathbb{E}\left[W_i(t_{ij}) + \epsilon_i(t_{ij})|\mathbf{b}_i\right]\right] \\
&= \mathbb{E}\left[\mathbb{V}ar\left[W_i(t_{ij})|\mathbf{b}_i\right] + \mathbb{V}ar\left[\epsilon_i(t_{ij})|\mathbf{b}_i\right]\right] + \mathbb{V}ar\left[\mathbb{E}\left[W_i(t_{ij})|\mathbf{b}_i\right] + \mathbb{E}\left[\epsilon_i(t_{ij})|\mathbf{b}_i\right]\right] \\
&= \mathbb{E}\left[\mathbb{V}ar\left[W_i(t_{ij})|\mathbf{b}_i\right] + Var\left[\epsilon_i(t_{ij})\right]\right] + \mathbb{V}ar\left[\mathbb{E}\left[W_i(t_{ij})|\mathbf{b}_i\right] + \mathbb{E}\left[\epsilon_i(t_{ij})\right]\right] \\
&= \mathbb{E}\left[\mathbb{V}ar\left[\mathbf{x}_i\boldsymbol{\beta} + \mathbf{f}(t_{ij})\mathbf{b}_i|\mathbf{b}_i\right] + \sigma_\epsilon^2\right] + \mathbb{V}ar\left[\mathbf{x}_i\boldsymbol{\beta} + \mathbf{f}(t_{ij})\mathbf{b}_i\right] \\
&= \mathbb{E}\left[\sigma_\epsilon^2\right] + \left[\mathbf{f}(t_{ij})\right]\mathbb{V}ar\left[\mathbf{b}_i\right]\left[\mathbf{f}(t_{ij})\right]^\top \\
&= \sigma_\epsilon^2 + \left[\mathbf{f}(t_{ij})\right]\Sigma_b\left[\mathbf{f}(t_{ij})\right]^\top.
\end{aligned}$$

If the Bernstein Polynomials of degree $m_L - 1$ is used to model the time-varying aspect of the longitudinal variable, these quantities are

$$
\begin{aligned}
\mathbb{E}[Y_i(t_{ij})] &= \mathbb{E}\left[\mathbb{E}[Y_i(t_{ij})|\boldsymbol{\xi}_i^{m_L-1}]\right] = \mathbb{E}[\mathbb{E}[W_i(t_{ij}) + \epsilon_i(t_{ij})|\boldsymbol{\xi}_i^{m_L-1}]] \\
&= \mathbb{E}[\mathbb{E}[W_i(t_{ij})|\boldsymbol{\xi}_i^{m_L-1}]] + \mathbb{E}[\mathbb{E}[\epsilon_i(t_{ij})|\boldsymbol{\xi}_i^{m_L-1}]] = \mathbb{E}[\mathbb{E}[W_i(t_{ij})|\boldsymbol{\xi}_i^{m_L-1}]] + \mathbb{E}[\epsilon_i(t_{ij})] \\
&= \mathbb{E}[\mathbb{E}[W_i(t_{ij}))|\boldsymbol{\xi}_i^{m_L-1}]] = \mathbb{E}\left[\mathbb{E}\left[\mathbf{x}_i\boldsymbol{\beta} + \mathbf{b}_{m_L-1}\left(\frac{t_{ij}}{T_{max}}\right)\boldsymbol{\xi}_i^{m_L-1}|\boldsymbol{\xi}_i^{m_L-1}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbf{x}_i\boldsymbol{\beta}|\boldsymbol{\xi}_i^{m_L-1}\right] + \mathbb{E}\left[\mathbf{b}_{m_L-1}\left(\frac{t_{ij}}{T_{max}}\right)\boldsymbol{\xi}_i^{m_L-1}|\boldsymbol{\xi}_i^{m_L-1}\right]\right] \\
&= \mathbb{E}\left[\mathbf{x}_i\boldsymbol{\beta} + \mathbf{b}_{m_L-1}\left(\frac{t_{ij}}{T_{max}}\right)\boldsymbol{\xi}_i^{m_L-1}\right] \\
&= \mathbb{E}\left[\mathbf{x}_i\boldsymbol{\beta}\right] + \mathbb{E}\left[\mathbf{b}_{m_L-1}\left(\frac{t_{ij}}{T_{max}}\right)\boldsymbol{\xi}_i^{m_L-1}\right] \\
&= \mathbf{x}_i\boldsymbol{\beta} + \mathbf{b}_{m_L-1}\left(\frac{t_{ij}}{T_{max}}\right)\mathbb{E}\left[\boldsymbol{\xi}_i^{m_L-1}\right] \\
&= \mathbf{x}_i\boldsymbol{\beta} + \mathbf{b}_{m_L-1}\left(\frac{t_{ij}}{T_{max}}\right)\boldsymbol{\mu}_\xi^{m_L-1}
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{V}ar[Y_i(t_{ij})] &= \mathbb{E}\left[\mathbb{V}ar\left[Y_i(t_{ij})|\boldsymbol{\xi}_i^{m_L-1}\right]\right] + \mathbb{V}ar\left[\mathbb{E}\left[Y_i(t_{ij})|\boldsymbol{\xi}_i^{m_L-1}\right]\right] \\
&= \mathbb{E}\left[\mathbb{V}ar\left[W_i(t_{ij}) + \epsilon_i(t_{ij})|\boldsymbol{\xi}_i^{m_L-1}\right]\right] + \mathbb{V}ar\left[\mathbb{E}\left[W_i(t_{ij}) + \epsilon_i(t_{ij})|\boldsymbol{\xi}_i^{m_L-1}\right]\right] \\
&= \mathbb{E}\left[\mathbb{V}ar\left[W_i(t_{ij})|\boldsymbol{\xi}_i^{m_L-1}\right] + Var\left[\epsilon_i(t_{ij})|\boldsymbol{\xi}_i^{m_L-1}\right]\right] + \mathbb{V}ar\left[\mathbb{E}\left[W_i(t_{ij})|\boldsymbol{\xi}_i^{m_L-1}\right] + \mathbb{E}\left[\epsilon_i(t_{ij})|\boldsymbol{\xi}_i^{m_L-1}\right]\right] \\
&= \mathbb{E}\left[\mathbb{V}ar\left[W_i(t_{ij})|\boldsymbol{\xi}_i^{m_L-1}\right] + Var\left[\epsilon_i(t_{ij})\right]\right] + \mathbb{V}ar\left[\mathbb{E}\left[W_i(t_{ij})|\boldsymbol{\xi}_i^{m_L-1}\right] + \mathbb{E}\left[\epsilon_i(t_{ij})\right]\right] \\
&= \mathbb{E}\left[\mathbb{V}ar\left[\mathbf{x}_i\boldsymbol{\beta} + \mathbf{b}_{m_L-1}\left(\frac{t_{ij}}{T_{max}}\right)\boldsymbol{\xi}_i^{m_L-1}|\boldsymbol{\xi}_i^{m_L-1}\right] + \sigma_\epsilon^2\right] + \mathbb{V}ar\left[\mathbf{x}_i\boldsymbol{\beta} + \mathbf{b}_{m_L-1}\left(\frac{t_{ij}}{T_{max}}\right)\boldsymbol{\xi}_i^{m_L-1}\right] \\
&= \mathbb{E}\left[\sigma_\epsilon^2\right] + \left[\mathbf{b}_{m_L-1}\left(\frac{t_{ij}}{T_{max}}\right)\right]\mathbb{V}ar\left[\boldsymbol{\xi}_i^{m_L-1}\right]\left[\mathbf{b}_{m_L-1}\left(\frac{t_{ij}}{T_{max}}\right)\right]^\top \\
&= \sigma_\epsilon^2 + \left[\mathbf{b}_{m_L-1}\left(\frac{t_{ij}}{T_{max}}\right)\right]\Sigma_{\xi^{m_L-1}}\left[\mathbf{b}_{m_L-1}\left(\frac{t_{ij}}{T_{max}}\right)\right]^\top \\
&= \sigma_\epsilon^2 + \sum_{k=1}^{m_L}\sum_{l=1}^{m_L} b_{k,m_L-1}\left(\frac{t_{ij}}{T_{max}}\right) b_{l,m_L-1}\left(\frac{t_{ij}}{T_{max}}\right)\sigma_{kl}.
\end{aligned}
$$

More information about this model is given in section 3.1.

Derivative of the BP approximation for the cumulative hazard function

The approximation of the Bernstein Polynomial of degree $m_S - 1$ for the baseline hazard function is obtained by taking the derivative of the approximation of the BP for the cumulative baseline hazard function. The model for H_0 and the calculations to obtain the derivative are given in this section.

The BP approximation with degree m_S for the baseline cumulative hazard function is

$$H_0(u) \approx \sum_{k=0}^{m_S} H_0\left(\frac{k}{m_S}T_{max}\right)\binom{m_S}{k}\left(\frac{u}{T_{max}}\right)^k\left(1-\frac{u}{T_{max}}\right)^{m_S-k}.$$

Then

$$\begin{aligned}
h_0(u) = H_0'(u) &\approx \sum_{k=0}^{m_S}\left[\frac{k}{T_{max}}H_0\left(\frac{k}{m_S}T_{max}\right)\binom{m_S}{k}\left(\frac{u}{T_{max}}\right)^{k-1}\left(1-\frac{u}{T_{max}}\right)^{m_S-k} - \right.\\
&\quad \left.\frac{(m_S-k)}{T_{max}}H_0\left(\frac{k}{m_S}T_{max}\right)\binom{m_S}{k}\left(\frac{u}{T_{max}}\right)^{k}\left(1-\frac{u}{T_{max}}\right)^{m_S-k-1}\right]\\
&= \sum_{k=0}^{m_S}\frac{1}{T_{max}}H_0\left(\frac{k}{m_S}T_{max}\right)\binom{m_S}{k}\left[k\left(\frac{u}{T_{max}}\right)^{k-1}\left(1-\frac{u}{T_{max}}\right)^{m_S-k} - (m_S-k)\left(\frac{u}{T_{max}}\right)^{k}\left(1-\frac{u}{T_{max}}\right)^{m_S-k-1}\right]\\
(\text{for } k=0) &= \frac{1}{T_{max}}H_0\left(\frac{0}{m_S}T_{max}\right)\binom{m_S}{0}\left[0\left(\frac{u}{T_{max}}\right)^{-1}\left(1-\frac{u}{T_{max}}\right)^{m_S} - (m_S-0)\left(\frac{u}{T_{max}}\right)^{0}\left(1-\frac{u}{T_{max}}\right)^{m_S-1}\right]\\
(\text{for } k=1) &= \frac{1}{T_{max}}H_0\left(\frac{1}{m_S}T_{max}\right)\binom{m_S}{1}\left[1\left(\frac{u}{T_{max}}\right)^{0}\left(1-\frac{u}{T_{max}}\right)^{m_S-1} - (m_S-1)\left(\frac{u}{T_{max}}\right)^{1}\left(1-\frac{u}{T_{max}}\right)^{m_S-2}\right]\\
(\text{for } k=2) &+ \frac{1}{T_{max}}H_0\left(\frac{2}{m_S}T_{max}\right)\binom{m_S}{2}\left[2\left(\frac{u}{T_{max}}\right)^{1}\left(1-\frac{u}{T_{max}}\right)^{m_S-2} - (m_S-2)\left(\frac{u}{T_{max}}\right)^{2}\left(1-\frac{u}{T_{max}}\right)^{m_S-3}\right]\\
(\text{for } k=3) &+ \frac{1}{T_{max}}H_0\left(\frac{3}{m_S}T_{max}\right)\binom{m_S}{3}\left[3\left(\frac{u}{T_{max}}\right)^{2}\left(1-\frac{u}{T_{max}}\right)^{m_S-3} - (m_S-3)\left(\frac{u}{T_{max}}\right)^{3}\left(1-\frac{u}{T_{max}}\right)^{m_S-4}\right]\\
&\vdots\\
(\text{for } k=m_S-2) &+ \frac{1}{T_{max}}H_0\left(\frac{m_S-2}{m_S}T_{max}\right)\binom{m_S}{m_S-2}\left[(m_S-2)\left(\frac{u}{T_{max}}\right)^{m_S-3}\left(1-\frac{u}{T_{max}}\right)^{2} - 2\left(\frac{u}{T_{max}}\right)^{m_S-2}\left(1-\frac{u}{T_{max}}\right)^{1}\right]\\
(\text{for } k=m_S-1) &+ \frac{1}{T_{max}}H_0\left(\frac{m_S-1}{m_S}T_{max}\right)\binom{m_S}{m_S-1}\left[(m_S-1)\left(\frac{u}{T_{max}}\right)^{m_S-2}\left(1-\frac{u}{T_{max}}\right)^{1} - 1\left(\frac{u}{T_{max}}\right)^{m_S-1}\left(1-\frac{u}{T_{max}}\right)^{0}\right]\\
(\text{for } k=m_S) &+ \frac{1}{T_{max}}H_0\left(\frac{m_S}{m_S}T_{max}\right)\binom{m_S}{m_S}\left[m_S\left(\frac{u}{T_{max}}\right)^{m_S-1}\left(1-\frac{u}{T_{max}}\right)^{m_S-m_S} - 0\left(\frac{u}{T_{max}}\right)^{m_S}\left(1-\frac{u}{T_{max}}\right)^{-1}\right].
\end{aligned}$$

The next step is to rearrange the expressions above by putting together the ones with the same exponents. In addition, note that $(m_S-(k-1))\binom{m_S}{k-1} = k\binom{m_S}{k}$, for $k = 1, 2, \ldots, m_S$. So, we have that

$$\begin{aligned}
h_0(u) &\approx \sum_{k=0}^{m_S}\frac{1}{T_{max}}H_0\left(\frac{k}{m_S}T_{max}\right)\binom{m_S}{k}\left[k\left(\frac{u}{T_{max}}\right)^{k-1}\left(1-\frac{u}{T_{max}}\right)^{m_S-k} - (m_S-k)\left(\frac{u}{T_{max}}\right)^{k}\left(1-\frac{u}{T_{max}}\right)^{m_S-k-1}\right]\\
&= \sum_{k=1}^{m_S}\frac{1}{T_{max}}\left[H_0\left(\frac{k}{m_S}T_{max}\right) - H_0\left(\frac{k-1}{m_S}T_{max}\right)\right]m_S\binom{m_S-1}{k-1}\left(\frac{u}{T_{max}}\right)^{k-1}\left(1-\frac{u}{T_{max}}\right)^{m_S-k}\\
&= \sum_{k=1}^{m_S}\frac{1}{T_{max}}\left[H_0\left(\frac{k}{m_S}T_{max}\right) - H_0\left(\frac{k-1}{m_S}T_{max}\right)\right]m_S f_{Binomial}\left(k-1; m_S-1, \frac{u}{T_{max}}\right)\\
&= \sum_{k=1}^{m_S}\frac{1}{T_{max}}\left[H_0\left(\frac{k}{m_S}T_{max}\right) - H_0\left(\frac{k-1}{m_S}T_{max}\right)\right]f_{Beta}\left(\frac{u}{T_{max}}; k, m_S-k+1\right).
\end{aligned}$$

Relationship between the summation of the components of $\boldsymbol{\gamma}_{m_S-1}$ and the baseline cumulative hazard function

As described in Section 3.2 (page 44), we have that $\gamma_k^{m_S-1} = H_0\left(\frac{k}{m_S}T_{max}\right) - H_0\left(\frac{k-1}{m_S}T_{max}\right)$, for $k = 1, 2, \ldots, m_S$ and $\gamma_0^{m_S-1} = 0$. Then,

$$\text{for } k=1, \quad \gamma_1^{m_S-1} = H_0\left(\frac{1}{m_S}T_{max}\right) - H_0\left(\frac{0}{m_S}T_{max}\right) \Longrightarrow H_0\left(\frac{1}{m_S}T_{max}\right) = \gamma_1^{m_S-1}$$

$$\text{for } k=2, \quad \gamma_2^{m_S-1} = H_0\left(\frac{2}{m_S}T_{max}\right) - H_0\left(\frac{1}{m_S}T_{max}\right) = H_0\left(\frac{2}{m_S}T_{max}\right) - \gamma_1^{m_S-1} \Longrightarrow H_0\left(\frac{2}{m_S}T_{max}\right) = \sum_{o=1}^{2}\gamma_o^{m_S-1}$$

$$\text{for } k=3, \quad \gamma_3^{m_S-1} = H_0\left(\frac{3}{m_S}T_{max}\right) - H_0\left(\frac{2}{m_S}T_{max}\right) = H_0\left(\frac{3}{m_S}T_{max}\right) - \gamma_2^{m_S-1} \Longrightarrow H_0\left(\frac{3}{m_S}T_{max}\right) = \sum_{o=1}^{3}\gamma_o^{m_S-1}$$

$$\vdots$$

$$\text{for } k=m_S, \quad \gamma_{m_S}^{m_S-1} = H_0\left(\frac{m_S}{m_S}T_{max}\right) - H_0\left(\frac{m_S-1}{m_S}T_{max}\right) = H_0\left(\frac{m_S}{m_S}T_{max}\right) - \gamma_{m_S-1}^{m_S-1} \Longrightarrow H_0\left(\frac{m_S}{m_S}T_{max}\right) = \sum_{o=1}^{m_S}\gamma_o^{m_S-1}.$$

Hence $H_0\left(\frac{k}{m_S}T_{max}\right) = \sum_{o=1}^{k}\gamma_k^{m_S-1}$.

Proof of the properties of the Bernstein Polynomials

The first property is the equivalency with the straight line. The BP approximation with degree m (arbitrary) is given by

$$\begin{aligned} f(t) &\approx \sum_{k=0}^{m} f\left(\frac{k}{m}\right)\binom{m}{k}t^k(1-t)^{m-k} = \sum_{k=0}^{m}\left(a\frac{k}{m}+b\right)\binom{m}{k}t^k(1-t)^{m-k} \\ &= \sum_{k=0}^{m} a\frac{k}{m}\binom{m}{k}t^k(1-t)^{m-k} + \sum_{k=0}^{m} b\binom{m}{k}t^k(1-t)^{m-k} \\ &= a\sum_{k=0}^{m}\frac{k}{m}\binom{m}{k}t^k(1-t)^{m-k} + b\sum_{k=0}^{m}\binom{m}{k}t^k(1-t)^{m-k}. \end{aligned}$$

Note that the terms in the second summation are probabilities from a $Binomial(m,t)$ distribution, for all possible number of successes. Then, these terms sum up to 1. So,

$$f(t) \approx a\sum_{k=0}^{m}\frac{k}{m}\binom{m}{k}t^k(1-t)^{m-k} + b = a\sum_{k=1}^{m}\frac{k}{m}\binom{m}{k}t^k(1-t)^{m-k} + b.$$

Using the result $\frac{k}{m}\binom{m}{k} = \binom{m-1}{k-1}$, we have that

$$\begin{aligned} f(t) &\approx a\sum_{k=1}^{m}\binom{m-1}{k-1}t^k(1-t)^{m-k}+b \\ &= at\sum_{k=1}^{m}\binom{m-1}{k-1}t^{k-1}(1-t)^{m-k}+b \\ &= at+b. \end{aligned}$$

The last result is true because the terms in the summation are probabilities of a $Binomial(m-1,t)$ distribution for all possible number of successes.

The demonstration of the degree elevation property starts by noticing that $1=(1-t)+1$. So, the Bernstein basis $b_{k,m}(t)=\binom{m}{k}t^k(1-t)^{m-k}$ is equal to

$$\begin{aligned} b_{k,m}(t) &= [(1-t)+1]\,b_{k,m}(t)=(1-t)\binom{m}{k}t^k(1-t)^{m-k}+t\binom{m}{k}t^k(1-t)^{m-k} \\ &= \binom{m}{k}t^k(1-t)^{(m+1)-k}+\binom{m}{k}t^{k+1}(1-t)^{m-k} \\ &= \frac{\binom{m}{k}}{\binom{m+1}{k}}\binom{m+1}{k}t^k(1-t)^{(m+1)-k}+\frac{\binom{m}{k}}{\binom{m+1}{k+1}}\binom{m+1}{k+1}t^{k+1}(1-t)^{m-k} \\ &= \left(1-\frac{k}{m+1}\right)b_{k,m+1}(t)+\left(\frac{k+1}{m+1}\right)b_{k+1,m+1}(t). \end{aligned}$$

As a result, we have for an arbitrary continuous function f that

$$\begin{aligned} f(t) &\approx \sum_{k=0}^{m} f\left(\frac{k}{m}\right)\binom{m}{k}t^k(1-t)^{m-k}=\sum_{k=0}^{m}\xi_k^m b_{k,m}(t) \\ &= \sum_{k=0}^{m}\xi_k^m\left[\left(1-\frac{k}{m+1}\right)b_{k,m+1}(t)+\left(\frac{k+1}{m+1}\right)b_{k+1,m+1}(t)\right] \\ &= \sum_{k=0}^{m}\xi_k^m\left(1-\frac{k}{m+1}\right)b_{k,m+1}(t)+\sum_{k=0}^{m}\xi_k^m\left(\frac{k+1}{m+1}\right)b_{k+1,m+1}(t) \\ (l=k+1) &= \sum_{k=0}^{m}\xi_k^m\left(1-\frac{k}{m+1}\right)b_{k,m+1}(t)+\sum_{l=1}^{m+1}\xi_{l-1}^m\left(\frac{l}{m+1}\right)b_{l,m+1}(t) \\ &= \sum_{k=0}^{m+1}\left[\xi_k^m\left(1-\frac{k}{m+1}\right)+\xi_{k-1}^m\left(\frac{k}{m+1}\right)\right]b_{k,m+1}(t) \\ &= \sum_{k=0}^{m+1}\tilde{\xi}_k^{m+1}b_{k,m+1}(t), \end{aligned}$$

where $\xi_0^{m+1}=\xi_0^m$ and $\xi_{m+1}^{m+1}=\xi_m^m$. For clarification, the term ξ_{k-1}^m does not make sense when $k=0$. However, this notation is not a concern, since the coefficient is being multiplied by zero.

Maximum of the Bernstein basis

The approximation for both the temporal behavior of the longitudinal variable and hazard function are given by a BP with degree $m-1$. Then, each of the m components of the Bernstein basis is $b_{k,m-1}\left(\frac{t}{T_{max}}\right) = \binom{m-1}{k-1}\left(\frac{t}{T_{max}}\right)^{k-1}\left(1-\frac{t}{T_{max}}\right)^{m-k}$ and its derivative is

$$b'_{k,m-1}\left(\frac{t}{T_{max}}\right) = \frac{(k-1)}{T_{max}}\binom{m-1}{k-1}\left(\frac{t}{T_{max}}\right)^{k-2}\left(1-\frac{t}{T_{max}}\right)^{m-k} - \frac{(m-k)}{T_{max}}\binom{m-1}{k-1}\left(\frac{t}{T_{max}}\right)^{k-1}\left(1-\frac{t}{T_{max}}\right)^{m-k-1}.$$

The point where this function reaches its maximum is a value t such that $b'_{k,m}\left(\frac{t}{T_{max}}\right) = 0$. Thus,

$$\begin{aligned}
& \frac{(k-1)}{T_{max}}\binom{m-1}{k-1}\left(\frac{t}{T_{max}}\right)^{k-2}\left(1-\frac{t}{T_{max}}\right)^{m-k} - \frac{(m-k)}{T_{max}}\binom{m-1}{k-1}\left(\frac{t}{T_{max}}\right)^{k-1}\left(1-\frac{t}{T_{max}}\right)^{m-k-1} = 0 \\
\Rightarrow\ & \frac{(k-1)}{T_{max}}\binom{m-1}{k-1}\left(\frac{t}{T_{max}}\right)^{k-2}\left(1-\frac{t}{T_{max}}\right)^{m-k} = \frac{(m-k)}{T_{max}}\binom{m-1}{k-1}\left(\frac{t}{T_{max}}\right)^{k-1}\left(1-\frac{t}{T_{max}}\right)^{m-k-1} \\
\Rightarrow\ & \left(\frac{t}{T_{max}}\right)^{-1}\left(1-\frac{t}{T_{max}}\right) = \frac{m-k}{k-1} \\
\Rightarrow\ & \frac{1-\frac{t}{T_{max}}}{\frac{t}{T_{max}}} = \frac{m-k}{k-1} \Rightarrow \left(\frac{t}{T_{max}}\right)^{-1} - 1 = \frac{m-k}{k-1} \\
\Rightarrow\ & \frac{t}{T_{max}} = \left(\frac{m-k}{k-1}+1\right)^{-1} = \left(\frac{m-1}{k-1}\right)^{-1} = \frac{k-1}{m-1}.
\end{aligned}$$

So, the maximum of the Bernstein basis is achieved when $t/T_{max} = (k-1)/(m-1) \Leftrightarrow t = ((k-1)/(m-1))T_{max}$.

Probability function and probability distribution function of M

The random variable M represents the minimum degree that is needed to capture a change in an interval $(U_{(1)}, U_{(2)})$. This variable is defined as $M = \left\lceil \max\left(\frac{1}{U_{(1)}}+1, \frac{2-U_{(2)}}{1-U_{(2)}}\right)\right\rceil$. So, we have that

$$\begin{aligned}
\mathbb{P}(M=m|(U_1,U_2)) &= \mathbb{P}\left(\left\lceil \max\left(\frac{1}{U_{(1)}}+1, \frac{2-U_{(2)}}{1-U_{(2)}}\right)\right\rceil = m|(U_1,U_2)\right) \\
&= \mathbb{P}\left(m-1 < \max\left(\frac{1}{U_{(1)}}+1, \frac{2-U_{(2)}}{1-U_{(2)}}\right) \leq m|(U_1,U_2)\right) \\
&= \mathbb{P}\left(\max\left(\frac{1}{U_{(1)}}+1, \frac{2-U_{(2)}}{1-U_{(2)}}\right) \leq m|(U_1,U_2)\right) - \mathbb{P}\left(\max\left(\frac{1}{U_{(1)}}+1, \frac{2-U_{(2)}}{1-U_{(2)}}\right) \leq m-1|(U_1,U_2)\right).
\end{aligned}$$

Now, note that

$$\begin{aligned}\left[\max\left(\frac{1}{U_{(1)}}+1,\frac{2-U_{(2)}}{1-U_{(2)}}\right)\leq m\right] &= \left[\frac{1}{U_{(1)}}+1\leq m,\frac{2-U_{(2)}}{1-U_{(2)}}\leq m\right]\\ &= \left[U_{(1)}\geq\frac{1}{m-1},U_{(2)}\leq\frac{m-2}{m-1}\right]\\ &= \left[\frac{1}{m-1}\leq U_1\leq\frac{m-2}{m-1},\frac{1}{m-1}\leq U_2\leq\frac{m-2}{m-1}\right].\end{aligned}$$

Consequently,

$$\left[\max\left(\frac{1}{U_{(1)}}+1,\frac{2-U_{(2)}}{1-U_{(2)}}\right)\leq m-1\right]=\left[\frac{1}{m-2}\leq U_1\leq\frac{m-3}{m-2},\frac{1}{m-2}\leq U_2\leq\frac{m-3}{m-2}\right].$$

In the case that U_1 and U_2 are equally distributed, we have that

$$\begin{aligned}\mathbb{P}(M=m|(U_1,U_2)) &= \mathbb{P}\left(\max\left(\frac{1}{U_{(1)}}+1,\frac{2-U_{(2)}}{1-U_{(2)}}\right)\leq m|(U_1,U_2)\right)-\mathbb{P}\left(\max\left(\frac{1}{U_{(1)}}+1,\frac{2-U_{(2)}}{1-U_{(2)}}\right)\leq m-1|(U_1,U_2)\right)\\ &= \left[\mathbb{P}\left(U_1\leq\frac{m-2}{m-1}\right)-\mathbb{P}\left(U_1<\frac{1}{m-1}\right)\right]^2-\left[\mathbb{P}\left(U_1\leq\frac{m-3}{m-2}\right)-\mathbb{P}\left(U_1<\frac{1}{m-2}\right)\right]^2.\end{aligned}$$

At last, considering $U_1\sim Beta(\theta_1,\theta_2)$ and $U_2\sim Beta(\theta_1,\theta_2)$

$$\mathbb{P}(M=m|(U_1,U_2)) = \left[F_{Beta}\left(\frac{m-2}{m-1};\theta_1,\theta_2\right)-F_{Beta}\left(\frac{1}{m-1};\theta_1,\theta_2\right)\right]^2-\left[F_{Beta}\left(\frac{m-3}{m-2};\theta_1,\theta_2\right)-F_{Beta}\left(\frac{1}{m-2};\theta_1,\theta_2\right)\right]^2.$$

Difference between two estimated curves

The difference between two approximated functions - one estimating the vector of coefficients directly ($f(t;m-1)$) and the other one obtained by using the degree elevation property ($\tilde{f}(t;m-1)$), is given by

$$
\begin{aligned}
D_{m-1} &= \int_0^1 (f(t;m-1)-\tilde{f}(t;m-1))^2 d(t/T_{max}) \\
&= \int_0^1 \left((\boldsymbol{\xi}_{m-1})^\top \mathbf{b}_{m-1}\left(\frac{t}{T_{max}}\right) - (\tilde{\boldsymbol{\xi}}_{m-1})^\top \mathbf{b}_{m-1}\left(\frac{t}{T_{max}}\right)\right)^2 d(t/T_{max}) \\
&= \int_0^1 \left[(\boldsymbol{\xi}_{m-1}^\top - \tilde{\boldsymbol{\xi}}_{m-1}^\top)\mathbf{b}_{m-1}\left(\frac{t}{T_{max}}\right)\right]^2 d(t/T_{max}) \\
&= \int_0^1 \left[(\boldsymbol{\xi}_{m-1} - \tilde{\boldsymbol{\xi}}_{m-1})^\top \mathbf{b}_{m-1}\left(\frac{t}{T_{max}}\right)\right]^2 d(t/T_{max}) \\
&= \int_0^1 \left[\sum_{k=1}^m \sum_{l=1}^m (\xi_k^{m-1} - \tilde{\xi}_k^{m-1})(\xi_l^{m-1} - \tilde{\xi}_l^{m-1}) b_{k,m-1}\left(\frac{t}{T_{max}}\right) b_{l,m-1}\left(\frac{t}{T_{max}}\right)\right] d(t/T_{max}) \\
&= \sum_{k=1}^m \sum_{l=1}^m (\xi_k^{m-1} - \tilde{\xi}_k^{m-1})(\xi_l^{m-1} - \tilde{\xi}_l^{m-1}) \int_0^1 \left[b_{k,m-1}\left(\frac{t}{T_{max}}\right) b_{l,m-1}\left(\frac{t}{T_{max}}\right)\right] d(t/T_{max}) \\
&= \sum_{k=1}^m \sum_{l=1}^m (\xi_k^{m-1} - \tilde{\xi}_k^{m-1})(\xi_l^{m-1} - \tilde{\xi}_l^{m-1}) \int_0^1 \binom{m-1}{k-1}\binom{m-1}{l-1}\left(\frac{t}{T_{max}}\right)^{k+l-2}\left(1-\frac{t}{T_{max}}\right)^{2m-k-l} d(t/T_{max}) \\
&= \sum_{k=1}^m \sum_{l=1}^m (\xi_k^{m-1} - \tilde{\xi}_k^{m-1})(\xi_l^{m-1} - \tilde{\xi}_l^{m-1}) \binom{m-1}{k-1}\binom{m-1}{l-1} \frac{\Gamma(k+l-1)\Gamma(2m-k-l+1)}{\Gamma(2m)} \\
&= \sum_{k=1}^m \sum_{l=1}^m (\xi_k^{m-1} - \tilde{\xi}_k^{m-1})(\xi_l^{m-1} - \tilde{\xi}_l^{m-1}) \frac{1}{(2m-1)} \frac{\binom{k+l-2}{k-1}\binom{2m-k-l}{m-k}}{\binom{2m-2}{m-1}} \\
&= \sum_{k=1}^m \sum_{l=1}^m (\xi_k^{m-1} - \tilde{\xi}_k^{m-1})(\xi_l^{m-1} - \tilde{\xi}_l^{m-1}) \frac{1}{(2m-1)} Hypergeometric(k-1;2m-2,k+l-2,m-1) \\
&= (\boldsymbol{\xi}_{m-1} - \tilde{\boldsymbol{\xi}}_{m-1})^\top \mathrm{A} (\boldsymbol{\xi}_{m-1} - \tilde{\boldsymbol{\xi}}_{m-1}),
\end{aligned}
$$

where A is an $m \times m$ matrix. Each component a_{kl} of this matrix is given by $a_{kl} = \frac{1}{(2m-1)} HyperG(k-1;2m-2,k+l-2,m-1)$, for $k = 1,2,\ldots,m$ and $l = 1,2,\ldots,m$. At last, $HyperG(k-1;2m-2,k+l-2,m-1)$ represents the probability of $k-1$ successes from an *Hypergeometric* distribution with parameters $(2m-2, k+l-2, m-1)$.

B Extra results of the simulation studies

Here, we will show details and extra results of the simulation studies. First, we give a brief description about the comparison measures we used to point to the best model, within the fitted ones. Then, we focus on the simulation study for the degree selection method we proposed. Next, we show results concerning the evaluation of the proposed modeling approach.

Comparison measures

In order to compare the fitted models and to be able to point out to the best one within this set, we used the DIC, the LPML and WAIC as comparison measures. Their formula are given below.

The DIC is based on the deviance measure, and it is described as

$$\begin{aligned} DIC &= \mathbb{E}\left[D(\boldsymbol{\Phi})|Data\right] + p_D \\ &= \mathbb{E}\left[D(\boldsymbol{\Phi})|Data\right] + \left\{\mathbb{E}\left[D(\boldsymbol{\Phi})|Data\right] - D(\mathbb{E}(\boldsymbol{\Phi})|Data)\right\} \end{aligned} \tag{B.1}$$

where $D(\boldsymbol{\Phi}) = -2\log(p(t|\boldsymbol{\Phi})$ is the deviance measure. In the DIC, the first component in the sum of Equation (B.1) accounts for the goodness of fit. Then, p_D concerns the model complexity (Gamerman and Lopes, 2006; Robert, 2007). This latter is also called the effective number of parameters. In practice, we can approximate the DIC in Equation (B.1) by

$$DIC \approx \frac{2}{S}\sum_{s=1}^{S} D(\boldsymbol{\Phi}^{(s)}) - D\left(\frac{1}{S}\sum_{s=1}^{S}\boldsymbol{\Phi}^{(s)}\right) \tag{B.2}$$

here, $\boldsymbol{\Phi}^{(s)}$ represents the s-th vector of the posterior sample, for $s = 1, 2, \ldots, S$.

In turn, the LPML measure is based on the Conditional Predictive Ordinate (CPO) (Ibrahim et al., 2001). The CPO is a quantity that measures how a specific observation influences the model. Thus, for a specific i-th observation, the CPO statistic is given by:

$$CPO_i = f(t_i|Data^{(-i)}) = \int_{\boldsymbol{\Phi}} f(t_i|\boldsymbol{\Phi}, Data^{(-i)})p(\boldsymbol{\Phi}|Data^{(-i)})d\boldsymbol{\Phi}, \tag{B.3}$$

where $Data^{(-i)}$ is the observed data excluding the i-th observation, $\boldsymbol{\Phi}$ is a general vector of parameters to be estimated. Since it is not possible to calculate expression B.3 analytically, an approximation is given by:

$$\widehat{CPO}_i = S\left\{\sum_{s=1}^{S}[f(t_i|\mathbf{\Phi}^l, D_{obs})]^{-1}\right\}^{-1},$$

where $s = 1, 2, \ldots, S$ represents the index of the posterior sample. Finally, the LPML measure can be obtained via:

$$LPML = \sum_{i=1}^{n} \log\left(CPO_i\right). \tag{B.4}$$

At last, the WAIC can be seen as a substitute to the DIC. This measure is an approximation to the following quantity:

$$\text{elpd} = \text{expected log pointwise predictive density for a new dataset} = \sum_{i=1}^{n} E_{f_i}[\log\left(p(\tilde{t}|Data)\right)],$$

where $\tilde{t}$ represents a new observation and $Data$ is the observed data.

The mentioned approximation is based on the posterior sample and it is given by:

$$\widehat{\text{elpd}}_{\text{waic}} = \widehat{\text{lpd}} - \hat{\text{p}}_{\text{waic}}, \tag{B.5}$$

where

$$\widehat{\text{lpd}} = \text{computed log pointwise predictive density} = \sum_{i=1}^{n} \log\left(\frac{1}{S}\sum_{s=1}^{S} f(t_i|\mathbf{\Phi}^s)\right),$$

likewise the notation to the LPML criteria, $s = 1, 2, \ldots, S$ is the index associated to the posterior sample, S is the size of the posterior sample and $\mathbf{\Phi}$ is the vector of estimated parameters.

In turn, $\hat{\text{p}}_{\text{waic}}$ is the estimated the effective number of parameters. It is also estimated using the posterior sample, in the following way:

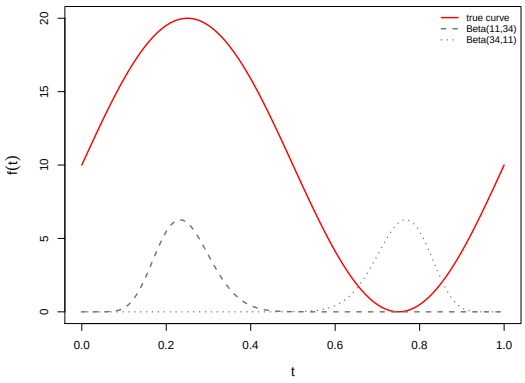
$$\hat{\text{p}}_{\text{waic}} = \sum_{i=1}^{n} V_{s=1}^{S}(\log\left(f(t_i|\mathbf{\Phi}^s)\right)), \tag{B.6}$$

where $V_{s=1}^{S} a_s = \dfrac{1}{S-1}\displaystyle\sum_{s=1}^{S}(a_s - \bar{a})^2$ is the sample variance.

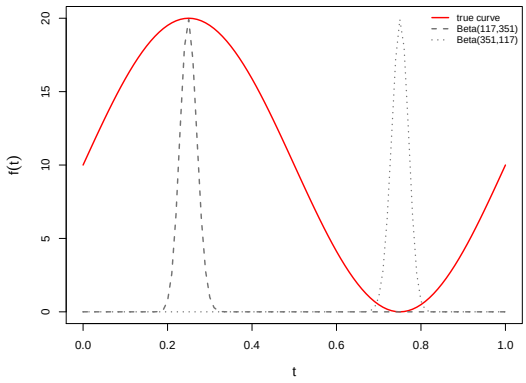
Next, we show extra results for both simulation studies.

Simulation Study 1: Degree selection

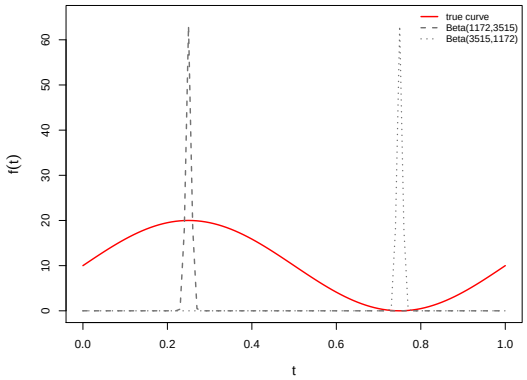
Figure 24 shows, in red, the true curve that was generated in the simulation study. The gray lines represent densities of the Beta distributions we chose to follow the routine we propose to define a minimum degree for the BP.



(a) $Beta(11, 34)$ and $Beta(34, 11)$.



(b) $Beta(117, 351)$ and $Beta(351, 117)$.



(c) $Beta(1172, 3515)$ and $Beta(3515, 1172)$.

Figure 24 – True curve $f(t)$ along with Beta densities that have high mass concentrated at the turning points.

We can observe that these densities are concentrated on the turning points of the true curve. The mean of these distributions are very close. On the other hand, as we can see, their variance diminishes from Figures 24a to 24c.

Simulation Study 2: Evaluation of the proposed model

In this section we will show details of the second simulation study. The focus of this other simulation study was to verify the performance of Bernstein Polynomials for both sub-models in the joint model framework. In addition, we compared this model to two others.

Details of data generation

Let $T \sim Gompertz(\alpha, \lambda)$, $\alpha \in \mathbb{R}$ is the shape parameter and $\lambda > 0$ is a scale parameter. Then, for $t > 0$, the survival functions are given by

$$\begin{aligned}
f_0(t) &= \lambda e^{at} \exp\left\{-\frac{\lambda}{a}\left(e^{at}-1\right)\right\},\\
F_0(t) &= 1-\exp\left\{-\frac{\lambda}{a}\left(e^{at}-1\right)\right\},\\
S_0(t) &= \exp\left\{-\frac{\lambda}{a}\left(e^{at}-1\right)\right\},\\
H_0(t) &= \frac{\lambda}{a}\left(e^{at}-1\right),\\
h_0(t) &= \lambda e^{at}.
\end{aligned}$$

Note that $h_0'(t) = \frac{d}{dt}h_0(t) = \alpha\lambda e^{at}$. Then, since $\lambda > 0$, if $\alpha > 0$, $h_0(.)$ is an increasing function of time. On the other hand, if $\alpha < 0$, then we have a decreasing function. In addition, considering the joint model structure described and detailed in Equations (4.2) and (4.3), the calculations necessary to generate joint longitudinal and survival data are given below.

The cumulative hazard function is

$$\begin{aligned}
H(u_i) &= \int_0^{u_i} h(w)dw\\
&= \exp\{\mathbf{z}_i\boldsymbol{\psi}\}\int_0^{u_i} \lambda e^{\alpha s}\exp\left\{\eta(\mathbf{x}_i\boldsymbol{\beta}+b_{0i}+b_{1i}s)\right\}ds\\
&= \exp\{\mathbf{z}_i\boldsymbol{\psi}\}\lambda\exp\{\eta(\mathbf{x}_i\boldsymbol{\beta}+b_{0i})\}\int_0^{u_i} e^{\alpha s}\exp\left\{\eta(b_{1i}s)\right\}ds\\
&= \exp\{\mathbf{z}_i\boldsymbol{\psi}\}\lambda\exp\{\eta(\mathbf{x}_i\boldsymbol{\beta}+b_{0i})\}\int_0^{u_i} \exp\left\{s\left[\alpha+\eta b_{1i}\right]\right\}ds\\
&= \exp\{\mathbf{z}_i\boldsymbol{\psi}\}\lambda\exp\{\eta\left(\mathbf{x}_i\boldsymbol{\beta}+b_{0i}\right)\}\left[\frac{1}{\alpha+\eta b_{1i}}\exp\left\{\left[\alpha+\eta b_{1i}\right]w\right\}\right]_{s=0}^{u_i}\\
&= \exp\{\mathbf{z}_i\boldsymbol{\psi}\}\lambda\exp\{\eta(\mathbf{x}_i\boldsymbol{\beta}+b_{0i})\}\left[\frac{1}{\alpha+\eta b_{1i}}\left(\exp\left\{\left[\alpha+\eta b_{1i}\right]u_i\right\}-1\right)\right].
\end{aligned}$$

The inverse function of the cumulative hazard function is given by

$$
\begin{aligned}
& H(u_i) = \exp\{\mathbf{z}_i\boldsymbol{\psi}\}\lambda\exp\{\eta(\mathbf{x}_i\boldsymbol{\beta}+b_{0i})\}\left[\frac{1}{\alpha+\eta b_{1i}}\left(\exp\left\{[\alpha+\eta b_{1i}]u_i\right\}-1\right)\right] \\
\Rightarrow \quad & \frac{\alpha+\eta b_{1i}}{\exp\{\mathbf{z}_i\boldsymbol{\psi}\}\lambda\exp\{\eta(\mathbf{x}_i\boldsymbol{\beta}+b_{0i})\}}H(u_i) = \exp\left\{[\alpha+\eta b_{1i}]u_i\right\}-1 \\
\Rightarrow \quad & \exp\left\{[\alpha+\eta b_{1i}]u_i\right\} = 1+\frac{\alpha+\eta b_{1i}}{\exp\{\mathbf{z}_i\boldsymbol{\psi}\}\lambda\exp\{\eta[\mathbf{x}_i\boldsymbol{\beta}+b_{0i}]\}}H(u_i) \\
\Rightarrow \quad & [\alpha+\eta b_{1i}]u_i = \log\left\{1+\frac{\alpha+\eta b_{1i}}{\exp\{\mathbf{z}_i\boldsymbol{\psi}\}\lambda\exp\{\eta[\mathbf{x}_i\boldsymbol{\beta}+b_{0i}]\}}H(u_i)\right\} \\
\Rightarrow \quad & u_i = \frac{1}{[\alpha+\eta b_{1i}]}\log\left\{1+\frac{\alpha+\eta b_{1i}}{\exp\{\mathbf{z}_i\boldsymbol{\psi}\}\lambda\exp\{\eta[\mathbf{x}_i\boldsymbol{\beta}+b_{0i}]\}}H(u_i)\right\}.
\end{aligned}
$$

Therefore,

$$
t = \frac{1}{[\alpha+\eta b_{1i}]}\log\left\{1+\frac{[\alpha+\eta b_{1i}](-\log(u^*))}{\exp\{\mathbf{z}_i\boldsymbol{\psi}\}\lambda\exp\{\eta[\mathbf{x}_i\boldsymbol{\beta}+b_{0i}]\}}\right\},
$$

where u^* is a value of random variable following a $Uniform(0,1)$ distribution.

Results of the case in which the estimation procedure neglects the correlation between measurements.

This section shows results of the simulation study for all fitted models when we did not use the Wishart distribution for the variance-covariance matrices. That is, when $\Sigma_b = \text{diag}(\sigma^2_{b_0}, \sigma^2_{b_1})$ and $\Sigma_\xi = \text{diag}(\sigma^2_{\xi_1}, \sigma^2_{\xi_2}, \ldots, \sigma^2_{\xi_{m_L}})$. In this case, the prior distributions were $\sigma_{b_l} \sim Gamma(1,1)$, for $l = 1, 2$, and $\sigma_{\xi_l} \sim Gamma(1,1)$, for $l = 1, 2, \ldots, m_L$.

Table 17 presents the coverage percentage for the main parameters based on the HPD interval. We highlight the lower coverage percentage for the variance/standard deviation parameters σ_{00}, σ_{11} and σ_ϵ. We can compare these results with those of Table 6. These outcomes show that to disregard the correlation between measurements lead to lower CP. A solution can be to model the variance-covariance matrix with a Wishart prior distribution. In addition, considering the approximation via BP, this interpretation is only valid for the standard deviation of the measurement error, since there are no related parameters for the standard deviations σ_{00} and σ_{11}. We also point out that there is no estimate for the covariance of the random effects $\sigma_{01} = \sigma_{10}$ for any of the fitted models, specifically because of to the modeling structure we applied here.

Table 18 shows results for the relative bias. We can note that, comparing to the RB of all other parameters, this measure for σ_{00} and σ_{11} are relatively high. Then, turning our attention to the standard deviation of the measurement error σ_ϵ, we can see that the RB is higher when using the BP to model the unknown functions. We can compare these values with those of Table 7. With this comparison, we can see that the RBs diminishes when we model the entire variance-covariance matrices Σ_b and Σ_ξ.

Table 17 – Coverage percentage based on HPD intervals for main the parameters. Here, the estimation procedure neglects the correlation between measurements.

	$\mathcal{M}^{N}_{Go}$	$\mathcal{M}^{N}_{We}$	$\mathcal{M}^{BP_5}_{BP_5}$	$\mathcal{M}^{BP_5}_{BP_{10}}$
μ_{b_0}	94.80	94.60	91.60	91.20
μ_{b_1}	96.40	94.80	-	-
β_1	94.40	94.80	85.00	84.60
β_2	96.00	96.40	87.40	86.20
σ_{00}	68.00	66.40	-	-
σ_{11}	70.40	71.00	-	-
σ_ϵ	88.20	86.60	38.40	37.60
ψ_1	97.00	51.20	97.60	97.60
ψ_2	94.40	93.80	94.00	95.20
η	96.40	62.00	95.00	95.40

Figures with complete results (including outliers)

Figures 25, 26, and 27 represents the figures shown in the simulation study chapter. Nonetheless, here they are shown including the outlier values. The panels in Figure 25 concern the parameters in the longitudinal sub-model. Then, we can relate these panels with Figures 9 to 12.

In turn, Figure 26 focuses on the parameters of the survival sub-models. These same figures without the outlier values are the Figures 13 to 15.

In Figure 27, the Panel 27a shows the comparison measures for all models. Thus, it includes these measures for the true model $\mathcal{M}^{N}_{Go}$. Panel 27b shows the comparison measures related to the true model, including the outliers. We can compare the outcomes in this panel with those in Figure 16.

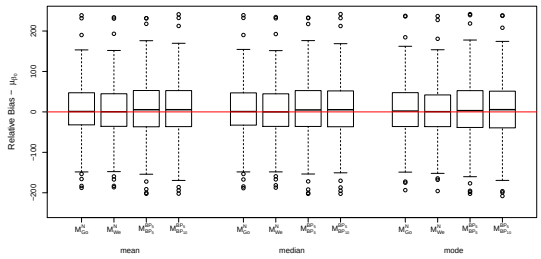
At last, Table 19 shows the frequency (and percentage) of how many times each model was chosen as the best one, for each of the three comparison measures.

Table 18 – Mean and standard deviations of the relative biases for the main parameters. Here, the estimation procedure neglects the correlation between measurements.

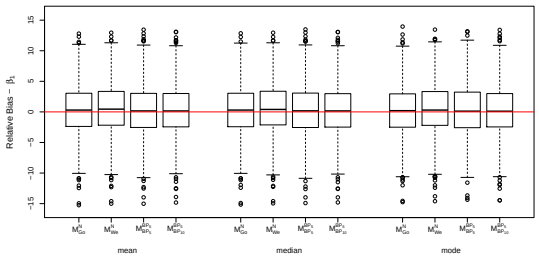
	Mean				Median				Mode			
	$\mathcal{M}_{Go}^{N}$	$\mathcal{M}_{We}^{N}$	$\mathcal{M}_{BP_5}^{BP_5}$	$\mathcal{M}_{BP_{10}}^{BP_5}$	$\mathcal{M}_{Go}^{N}$	$\mathcal{M}_{We}^{N}$	$\mathcal{M}_{BP_5}^{BP_5}$	$\mathcal{M}_{BP_{10}}^{BP_5}$	$\mathcal{M}_{Go}^{N}$	$\mathcal{M}_{We}^{N}$	$\mathcal{M}_{BP_5}^{BP_5}$	$\mathcal{M}_{BP_{10}}^{BP_5}$
μ_{b_0}	2.64 (63.41)	1.07 (63.51)	1.87 (71.57)	1.62 (71.64)	2.10 (63.41)	0.98 (63.61)	1.95 (71.56)	1.82 (71.58)	-4.93 (63.41)	0.96 (64.64)	2.04 (72.00)	2.71 (72.59)
μ_{b_1}	-1.12 (6.37)	-2.76 (6.41)	- -	- -	-1.15 (6.37)	-2.80 (6.42)	- -	- -	-1.17 (6.42)	-2.83 (6.53)	- -	- -
β_1	0.00 (4.51)	0.02 (4.52)	-1.97 (4.98)	-1.94 (4.98)	0.05 (4.51)	0.01 (4.53)	-1.98 (4.98)	-1.94 (4.97)	-0.58 (4.51)	0.02 (4.56)	-2.00 (5.03)	-1.90 (5.03)
β_2	0.05 (34.20)	0.07 (34.15)	0.81 (38.20)	0.99 (38.29)	-0.27 (34.20)	0.06 (34.16)	0.74 (38.22)	0.90 (38.31)	-1.59 (34.20)	-0.14 (34.49)	0.56 (38.54)	0.92 (38.42)
σ_{00}	20.80 (13.34)	21.32 (13.33)	- -	- -	20.19 (13.28)	20.71 (13.28)	- -	- -	19.03 (13.26)	19.70 (13.21)	- -	- -
σ_{11}	21.11 (15.12)	21.05 (15.09)	- -	- -	20.27 (15.04)	20.20 (15.02)	- -	- -	18.74 (14.98)	18.66 (14.91)	- -	- -
σ_{ϵ}	-1.40 (2.20)	-1.52 (2.20)	-6.84 (3.26)	-6.80 (3.26)	-1.33 (2.20)	-1.56 (2.19)	-6.90 (3.26)	-6.85 (3.27)	-1.19 (2.20)	-1.64 (2.20)	-7.00 (3.30)	-6.94 (3.32)
ψ_1	0.75 (6.21)	-11.73 (5.95)	1.30 (6.55)	1.01 (6.48)	0.35 (6.21)	-11.81 (5.94)	1.20 (6.54)	0.90 (6.48)	-0.61 (6.21)	-12.00 (6.00)	1.00 (6.57)	0.69 (6.52)
ψ_2	1.43 (27.03)	13.80 (24.59)	0.20 (27.22)	-1.07 (27.05)	1.76 (27.03)	13.88 (24.57)	0.32 (27.21)	-0.96 (27.12)	3.06 (27.03)	14.00 (24.82)	0.56 (27.49)	-0.87 (27.74)
η	0.45 (12.53)	-22.23 (12.35)	2.54 (13.77)	2.49 (13.64)	-0.28 (12.53)	-22.35 (12.32)	2.34 (13.73)	2.28 (13.63)	-2.03 (12.53)	-22.51 (12.36)	1.90 (13.91)	1.88 (13.83)

Table 19 – Frequency and percentage of the times in which each model was chosen as the best one.

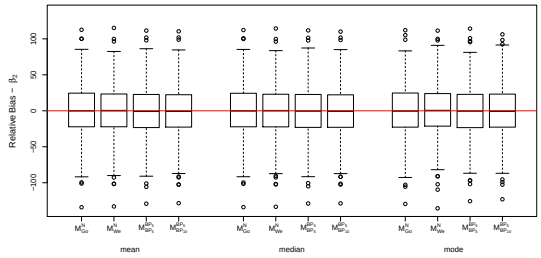
	$\mathcal{M}_{Go}^{N}$	$\mathcal{M}_{We}^{N}$	$\mathcal{M}_{BP_5}^{BP_5}$	$\mathcal{M}_{BP_{10}}^{BP_5}$
DIC	337 (67.40%)	36 (7.20%)	81 (16.20%)	46 (9.20%)
LPML	446 (89.20%)	2 (0.40%)	37 (7.40%)	15 (3.00%)
WAIC	448 (89.60%)	2 (0.40%)	36 (7.20%)	14 (2.80%)



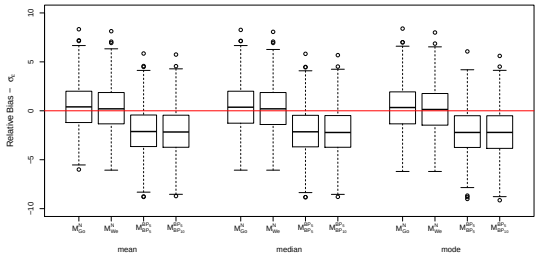
(a) μ_{b_0}.



(b) β_1.

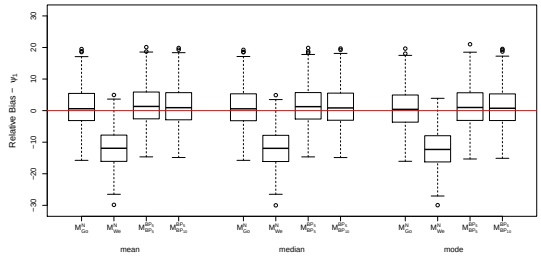


(c) β_2.

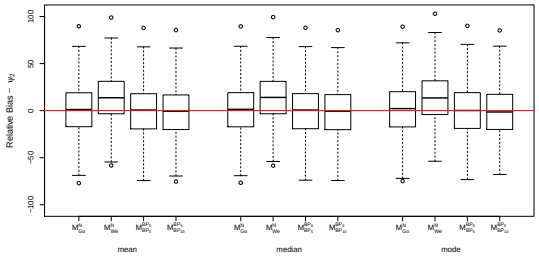


(d) σ_ϵ.

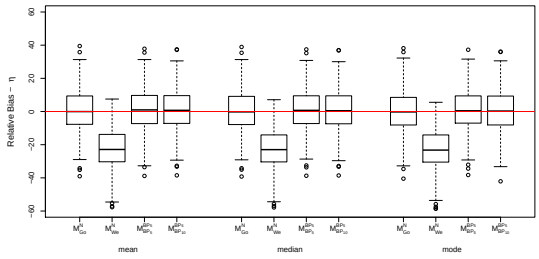
Figure 25 – Comparison of the relative biases of the parameters related to the *longitudinal* sub-model based on the posterior mean, median and mode and comparing each modeling approach.



(a) ψ_1.

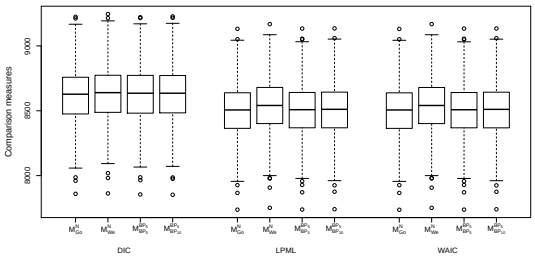


(b) ψ_2.

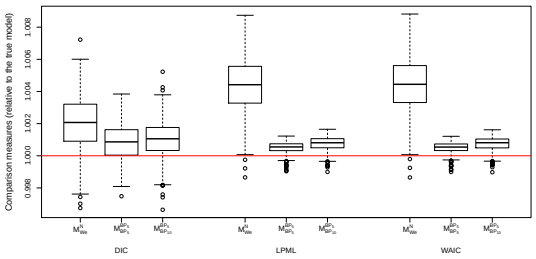


(c) η.

Figure 26 – Comparison of the relative biases of the parameters related to the *survival* sub-model based on the posterior mean, median and mode and comparing each modeling approach.



(a) Originally scaled comparison measured for all fitted models.



(b) Figure 16 with outliers.

Figure 27 – Comparison measures for the fitted models.

C Extra results of the application

This chapter shows an extra result of the application. Figure 28 shows some of the estimated models for the data set and the trajectory of all subjects under study. Thus, this figure is the same as Figure 21c but including the trajectories of subjects and the observed mean curve, as it was done in Figure 19b. This figure was made to allow a comparison with the observed longitudinal values for each person, the observed mean and the estimated mean curve for the selected models.

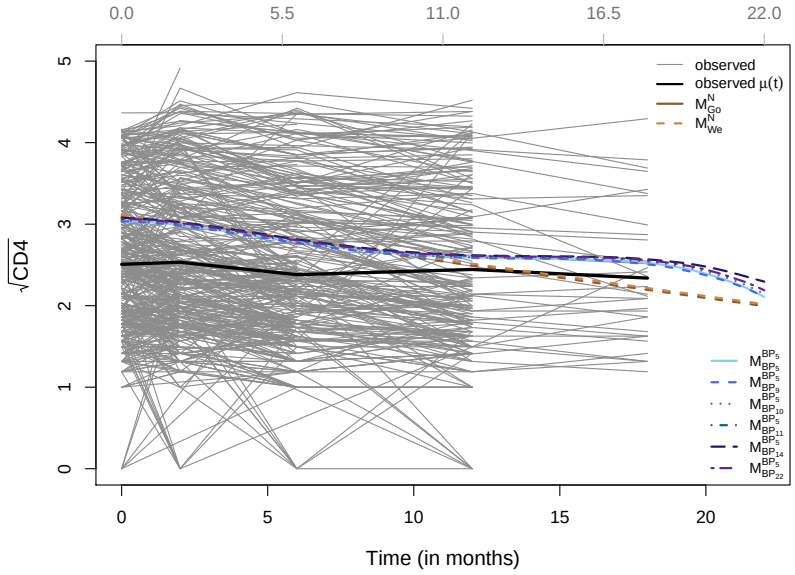


Figure 28 – Comparison between estimated overall mean curve along with the trajectory of all subjects.

Bibliography

Abrams, D. I., Goldman, A. I., Launer, C., Korvick, J. A., Neaton, J. D., Crane, L. R., Grodesky, M., Wakefield, S., Muth, K., Kornegay, S., Cohn, D. L., Harris, A., Luskin-Hawk, R., Markowitz, N., Sampson, J. H., Thompson, M., Deyton, L. and the Terry Beirn Community Programs for Clinical Research on AIDS (1994) A comparative trial of didanosine or zalcitabine after treatment with zidovudine in patients with human immunodeficiency virus infection. *New England Journal of Medicine*, **330**, 657–662.

Arjas, E. and Gasbarra, D. (1994) Nonparametric Bayesian inference from right censored survival data, using the Gibbs sampler. *Statistica Sinica*, **4**, 505–524.

Arnholt, A. T. and Evans, B. (2017) *BSDA: Basic Statistics and Data Analysis*. URLhttps://CRAN.R-project.org/package=BSDA. R package version 1.2.0.

Austin, P. C. (2012) Generating survival times to simulate Cox proportional hazards models with time-varying covariates. *Statistics in Medicine*, **31**, 3946–3958.

Babu, G. J., Canty, A. J. and Chaubey, Y. P. (2002) Application of Bernstein polynomials for smooth estimation of a distribution and density function. *Journal of Statistical Planning and Inference*, **105**, 377 – 392.

Bartle, R. G. and Sherbert, D. R. (2011) *Introduction to Real Analysis*. Wiley, 4 edn.

Bender, R., Augustin, T. and Blettner, M. (2005) Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, **24**, 1713–1723.

Bennett, S. (1983) Analysis of survival data by the proportional odds model. *Statistics in Medicine*, **2**, 273–277.

Bernstein, S. N. (1912) Démonstration du théorème de Weiertrass fondée sur le calcul des probabilités. *Kharkov Mathematical Society*, **13**.

Bertrand, A., Keilegom, I. V. and Legrand, C. (2019) Flexible parametric approach to classical measurement error variance estimation without auxiliary data. *Biometrics*, **75**, 297–307.

Birnbaum, Z. W. and Saunders, S. C. (1969) A new family of life distributions. *Journal of Applied Probability*, **6**, 319–327.

Borchers, H. W. (2019) *pracma: Practical Numerical Math Functions*. URLhttps://CRAN.R-project.org/package=pracma. R package version 2.2.9.

Brilleman, S., Crowther, M., Moreno-Betancur, M., Novik, J. B. and Wolfe, R. (2017) Joint longitudinal and time-to-event models via Stan.

Brown, B. M. and Chen, S. X. (1999) Beta-Bernstein smoothing for regression curves with compact support. *Scandinavian Journal of Statistics*, **26**, 47–59.

Brown, E. R. and Ibrahim, J. G. (2003) A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics*, **59**, 221–228.

Carnicer, J. M. and Peña, J. M. (1993) Shape preserving representations and optimality of the Bernstein basis. *Advances in Computational Mathematics*, **1**, 173–196.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. and Riddell, A. (2017) Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, **76**, 1–32. URLhttps://www.jstatsoft.org/v076/i01.

Chang, I.-S., Chien, L.-C., Hsiung, C. A., Wen, C.-C. and Wu, Y.-J. (2007) Shape restricted regression with random Bernstein polynomials. *Lecture Notes-Monograph Series*, **54**, 187–202.

Chang, I.-S., Hsiung, C. A., Yuh-Jennwu and Yang, C.-C. (2005) Bayesian survival analysis using Bernstein polynomials. *Scandinavian Journal of Statistics*, **32**, 447–466.

Chen, Y., Hanson, T. and Zhang, J. (2014) Accelerated hazards model based on parametric families generalized with Bernstein polynomials. *Biometrics*, **70**, 192–201.

Colosimo, E. A. and Giolo, S. R. (2006) *Análise de sobrevivência aplicada*. ABE - Projeto Fisher. Edgard Blücher.

Cox, D. R. (1972) Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, **34**, 187–220.

Crowther, M. J., Abrams, K. R. and Lambert, P. C. (2012) Flexible parametric joint modelling of longitudinal and survival data. *Statistics in Medicine*, **31**, 4456–4471.

Crowther, M. J. and Lambert, P. C. (2013) Simulating biologically plausible complex survival data. *Statistics in Medicine*, **32**, 4118–4134.

Curtis, S. M. and Ghosh, S. K. (2011) A variable selection approach to monotonic regression with Bernstein polynomials. *Journal of Applied Statistics*, **38**, 961–976.

Farouki, R. T. (2012) The Bernstein polynomial basis: A centennial retrospective. *Computer Aided Geometric Design*, **29**, 379 – 419.

Faucett, C. L. and Thomas, D. C. (1996) Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine*, **15**, 1663–1685.

de Figueiredo, D. G. (1996) *Análise I*. Rio de Janeiro - RJ: LTC - Livros Técnicos e Científicos Editora S. A., 2nd edn.

Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2012) *Applied Longitudinal Analysis*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series. John Wiley & Sons, 2nd edn.

Gamerman, D. and Lopes, H. F. (2006) *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall/CRC Texts in Statistical Science. New York: Taylor & Francis, 2nd edn.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003) *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science. New York: Taylor & Francis, 2 edn.

Gibbons, J. D. and Chakraborti, S. (2003) *Nonparametric Statistical Inference*. Statistics: Textbooks & Monograph. Chapman and Hall/CRC, 4th edn.

Guan, Z. (2016) Efficient and robust density estimation using Bernstein type polynomials. *Journal of Nonparametric Statistics*, **28**, 250–271.

Guo, X. and Carlin, B. P. (2004) Separate and joint modeling of longitudinal and event time data using standard computer packages. *The American Statistician*, **58**, 16–24.

Harville, D. A. (1977) Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, **72**, 320–338.

Ibrahim, J. G., Chen, M.-H. and Sinha, D. (2001) *Bayesian Survival Analysis*. Springer Series in Statistics. New York: Springer-Verlag, 1st edn.

Ibrahim, J. G., Chu, H. and Chen, L. M. (2010) Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology*, **28**, 2796–2801.

Kahaner, D., Moler, C. and Nash, S. (1989) *Numerical Methods and Software*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Kaplan, E. L. and Meier, P. (1958) Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457–481.

Kirmani, S. N. U. A. and Gupta, R. C. (2001) On the proportional odds model in survival analysis. *Annals of the Institute of Statistical Mathematics*, **53**, 203–216.

Klein, J. P., van Houwelingen, H. C., Ibrahim, J. G. and Scheike, T. H. (2013) *Handbook of Survival Analysis*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. New York: Chapman and Hall/CRC, 1st edn.

Klein, J. P. and Moeschberger, M. L. (2003) *Survival Analysis: Techniques for Censored and Truncated Data*. Statistics for Biology and Health. Springer-Verlag, New York, 2nd edn.

Kottas, A. (2006) Dirichlet process mixtures of Beta distributions, with applications to density and intensity estimation. In *In Proceedings of the Workshop on Learning with Nonparametric Bayesian Methods*.

Kuller, R. G. (1964) Coin tossing, probability, and the Weierstrass approximation theorem. *Mathematics Magazine*, **37**, 262–265.

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

Lawless, J. F. (2003) *Statistical Models and Methods for Lifetime Data*. Wiley Series in Probability and Statistics. Wiley-Interscience, 2 edn.

Lorentz, G. G. (1986) *Bernstein Polynomials*, vol. 323 of *AMS Chelsea Publishing*. American Mathematical Society.

Osman, M. and Ghosh, S. K. (2012) Nonparametric regression models for right-censored data using Bernstein polynomials. *Computational Statistics & Data Analysis*, **56**, 559 – 573.

Petrone, S. (1999a) Bayesian density estimation using Bernstein polynomials. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, **27**, 105–126.

— (1999b) Random Bernstein polynomials. *Scandinavian Journal of Statistics*, **26**, 373–393.

R Core Team (2019) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URLhttps://www.R-project.org/.

Rizopoulos, D. (2010) JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, **35**, 1–33. URLhttp://www.jstatsoft.org/v35/i09/.

— (2012) *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. Chapman & Hall/CRC Biostatistics Series. New York: Chapman and Hall/CRC, 1 edn.

Rizopoulos, D., Verbeke, G. and Molenberghs, G. (2010) Multiple-imputation-based residuals and diagnostic plots for joint models of longitudinal and survival outcomes. *Biometrics*, **66**, 20–29.

Robert, C. (2007) *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer Texts in Statistics. New York: Springer-Verlag, 2nd edn.

Rodrigues, J., Cancho, V. G. and de Castro, M. (2008) Teoria unificada de análise de sobrevivência.

Spiegelhalter, D. J., Best, N. G. and Carlin, B. P. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*.

Stan Development Team (2018) Stan modeling language users guide and reference manual, version 2.18.0. URLhttp://mc-stan.org/.

— (2019) RStan: the R interface to Stan. URLhttp://mc-stan.org/. R package version 2.19.2.

Tsiatis, A. A., Dafni, U., DeGruttola, V., Propert, K. J., Strawderman, R. L. and Wulfsohn, M. (1992) *The Relationship of CD4 Counts over Time to Survival in Patients with AIDS: Is CD4 a Good Surrogate Marker?*, 256–274. Boston, MA: Birkhäuser Boston.

Tsiatis, A. A. and Davidian, M. (2004) Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica*, **14**, 809–834.

Tsiatis, A. A., Degruttola, V. and Wulfsohn, M. S. (1995) Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*, **90**, 27–37.

Vehtari, A. and Gelman, A. (2014) WAIC and cross-validation in Stan ∗.

Vitale, R. A. (1975) A Bernstein polynomial approach to density function estimation. In *Statistical Inference and Related Topics* (ed. M. L. Puri), 87 – 99. Academic Press.

Wang, J. and Ghosh, S. K. (2012) Shape restricted nonparametric regression with Bernstein polynomials. *Computational Statistics & Data Analysis*, **56**, 2729 – 2741.

Wang, L. and Ghosh, S. K. (2013) Nonparametric models for longitudinal data using Bernstein polynomial sieve. *Tech. rep.*, Department of Statistics, North Carolina State University. URLhttps://repository.lib.ncsu.edu/bitstream/handle/1840.4/8245/mimeo2651_Wang.pdf?sequence=1&isAllowed=y.

Watanabe, S. (2010) Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, **11**, 3571–3594.

Wu, L., Liu, W., Yi, G. Y. and Huang, Y. (2012) Analysis of longitudinal and survival data: Joint modeling, inference methods, and issues. *Journal of Probability and Statistics*, **2012**, 1–17.

Wu, M. C. and Bailey, K. (1988) Analysing changes in the presence of informative right censoring caused by death and withdrawal. *Statistics in Medicine*, **7**, 337–346.

Wulfsohn, M. S. and Tsiatis, A. A. (1997) A joint model for survival and longitudinal data measured with error. *Biometrics*, **53**, 330–339.

Xu, J. and Zeger, S. L. (2001) Joint analysis of longitudinal data comprising repeated measures and times to events. *Journal of the Royal Statistical Society, Series C*, **50**, 375–387.

Zhou, H. and Hanson, T. (2018) A unified framework for fitting Bayesian semiparametric models to arbitrarily censored survival data, including spatially referenced data. *Journal of the American Statistical Association*, **113**, 571–581.

Zhou, Q., Hu, T. and Sun, J. (2017) A sieve semiparametric maximum likelihood approach for regression analysis of bivariate interval-censored failure time data. *Journal of the American Statistical Association*, **112**, 664–672.

Zhu, H., Ibrahim, J. G., Chi, Y.-Y. and Tang, N. (2012) Bayesian influence measures for joint models for longitudinal and survival data. *Biometrics*, **68**, 954–964.