UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

PAULO CERQUEIRA DOS SANTOS JUNIOR

MODELOS SEMIPARAMÉTRICOS PARA DADOS DE SOBREVIVÊNCIA COM CENSURA INTERVALAR

Belo Horizonte
Novembro de 2016

PAULO CERQUEIRA DOS SANTOS JUNIOR

MODELOS SEMIPARAMÉTRICOS PARA DADOS DE SOBREVIVÊNCIA COM CENSURA INTERVALAR

Tese de doutorado apresentada ao Programa de Pós-graduação em Estatística do Departamento de Estatística da Universidade Federal de Minas Gerais como parte dos requisitos para a obtenção do grau de Doutor em Estatística.

Orientador: Prof. Dr. Fábio Nogueira Demarqui

Coorientadora: Profa. Dra. Lourdes Contreras Montenegro

Belo Horizonte
2016

MODELOS SEMIPARAMÉTRICOS PARA DADOS DE SOBREVIVÊNCIA COM CENSURA INTERVALAR

Esta tese trata-se da versão original
do aluno Paulo Cerqueira dos Santos Junior.

MODELOS SEMIPARAMÉTRICOS PARA DADOS DE SOBREVIVÊNCIA COM CENSURA INTERVALAR

Esta tese contém as correções e alterações
sugeridas pela Comissão Julgadora durante a defesa
realizada por Paulo Cerqueira dos Santos Junior em 25/11/2016.
O original encontra-se disponível no Departamento de
de Estatística da Universidade Federal de Minas Gerais.

Comissão Julgadora:

- Prof. Dr. Fábio Nogueira Demarqui (Orientador) - UFMG
- Profa. Dra. Lourdes Contreras Montenegro - UFMG (Coorientadora)
- Profa. Dra. Rosangela Helena Loschi - UFMG
- Prof. Dr. Vinicius Diniz Mayrink - UFMG
- Prof. Dr. Mário Andrade de Castro - ICMC - USP
- Prof. Dr. Vicente Cancho Garibay - ICMC - USP

Para o meu filho, Bento.

Agradecimentos

⋆ Agradeço ao meu Deus e Nossa Senhora pelas oportunidades que são colocadas em minha vida, por seu imenso amor e por sempre abençoar-me em meus passos;

⋆ Ao meu Orientador Fábio Nogueira Demarqui pela oportunidade de ser seu orientando de doutorado. Obrigado pela paciência e por compartilhar parte de sua experiência de trabalho. Valeu Fábio;

⋆ À minha Coorientadora Lourdes Montenegro por aceitar me orientar no início neste trabalho, dessa forma, hoje não teríamos nada. Obrigado Lourdes;

⋆ À minha esposa Fabrícia e meu filho Bento, minhas fontes de inspiração e motivação. Agradeço a Deus por tê-los em minha vida. Te amo;

⋆ Aos meus pais, Paulo e Marizete, por todo carinho, amor, incentivos e suporte para mais uma etapa em minha vida;

⋆ Ao Progama de Medicina Tropical pela oportunidade de bolsa e pela nova experiência de trabalho na medicina tropical.

⋆ Ao departamento de estatística e ao programa de doutorado em estatística pela oportunidade de bolsa com incentivos para a pesquisa e docência através da bolsa reúne.

⋆ Aos membros avaliadores deste trabalho de teste: Profa. Rosangela Loschi, Prof. Vinicius Mayrink, Prof. Mário de Castro e Prof. Vicente Garibay, pelas ótimas sugestões e comentários.

⋆ Agradeço pela minha família formada em BH por esses seis anos, Rodrigo, Luis, Wecsley, Nívea, Lívia, Larissa, Juliana, Juliane, Márcio, Denise, por todos os momentos maravilhosos que pude desfrutar em minha estadia em BH. Agradeço também aos amigos, pelos diversos conselhos e orientações que serviram de alicerce para o desenvolvimento deste trabalho.

⋆ Quero agradecer à Rogéria pela paciência e delicadeza em me ajudar nos momentos de agonia;

⋆ À CAPES pelo apoio financeiro;

E nunca te esquece, se as coisas ainda não deram certo:
Trabalha, trabalha e trabalha!
Papai.

Resumo

Em análise de sobrevivência, dizemos que a censura é intervalar quando sabe-se somente que o tempo de sobrevivência dos indivíduos em análise pertence a um intervalo de tempo. Assim, esta tese de doutorado tem como objetivo propor extensões ao modelo exponencial por partes com grade aleatória, via estrutura de agrupamento do modelo partição produto, sob uma abordagem dinâmica, para dados de sobrevivência sujeitos à censura intervalar. Propomos modelos para dois tipos de população: sem e com fração de curados. Para dados sem fração de curados, apresentamos três propostas de modelagem. A primeira, representando o MEP com grade aleatória com efeito da covariável fixo no tempo, e as seguintes representando extensões de dois modelos dinâmicos, um com grade fixa e outro com grade aleatória, em que ambos permitem que o efeito da covariável varie no tempo. Para este tipo de população, ilustramos os modelos propostos, e modelos disponíveis na literatura, com dados de câncer de mama, avaliando o tempo até a deterioração da mama das pacientes. Como resultado principal, observamos que o modelo dinâmico com grade fixa e efeito variando no tempo proposto apresentou os melhores resultados quando comparado aos demais. Em situações em que há presença de uma fração de curados na população, apresentamos duas propostas de modelos com fração de cura. A primeira é um MEP dinâmico com grade fixa e a segunda sendo a versão com grade aleatória via estrutura de agrupamento do modelo partição produto, construídos com base no modelo de tempos de promoção. Neste cenário, ilustramos os modelos propostos utilizando dados de tempos de infecção por HIV-1 em pacientes hemofílicos. Os resultados mostraram que o modelo com fração de cura dinâmico com grade aleatória, quando comparado com o modelo dinâmico com grade fixa, apresentou a melhor adequação aos dados.

Palavras-chave: Censura intervalar, Fração de cura, Modelagem dinâmica, MEP com grade aleatória.

Abstract

In survival analysis we say that the data is subejct to interval-censored when it is known that the survival time of the individuals is in a interval time. Thus, this thesis aims to propose extensions to the piecewise exponential model with random grid, via the product partition model clustering structure under a dynamic approach, to survival data subject to interval-censored. We propose models for two types of populations: with and without a cured fraction. For population without a cured fraction, we present three modeling proposal. The first one representing the PEM with random time grid, with time-independent coefficients. In addition two dynamic models representing extensions, one with fixed time grid and another with random time grid, both allowing time-dependent coefficients. For this type of population, we illustrate the proposed models, and those available in the literature, using a breast cancer data, analysing the deterioration time of the patients. The main result shows that the proposed dynamic model with fixed time grid and time-independent coefficient has the best results when compared to the others. For situations with cured fraction in the population, we present two proposals of cure rate models. The first one, is the dynamic MEP with fixed time grid, and the second being the random time grid version, using the product partition model clustering structure, built based on the promotion time model. In this scenario, we illustrate an proposed models, using the infection time data. The results show that the dynamic cure rate model with a random time grid, showed the best fit to the data, when compared with the dynamic cure rate model with a fixed grid time.

Keywords: Interval-censored, Cure rate model, Dinamic approach, Random grid PEM.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Análise de sobrevivência é a expressão utilizada para designar a análise estatística de dados quando a variável resposta em estudo representa o tempo desde um instante inicial até a ocorrência de determinado evento de interesse. O evento de interesse pode ser, por exemplo, a morte de um paciente, a cura ou recidiva de uma doença ou a ocorrência de uma falha em um equipamento.

O estado da arte da análise de sobrevivência é o produto de um longo processo que sofreu um grande crescimento nos últimos 50 anos. De acordo com Colosimo & Giolo (2006), a análise de sobrevivência é uma das áreas da estatística que mais cresceu nas últimas décadas, impulsionada pelo aperfeiçoamento de procedimentos estatísticos e computacionais observados neste período.

Uma característica peculiar dos dados de sobrevivência é a presença de censura, caracterizada pela observação incompleta da variável de interesse para alguns indivíduos do estudo, seja por iniciativa dos mesmos em sair do estudo, perda de contato durante o acompanhamento, término do estudo sem a ocorrência do evento de interesse, etc. A censura pode ser classificada como:

- Censura à direita: o tempo de ocorrência do evento de interesse é maior que o tempo registrado.

- Censura à esquerda: o tempo registrado é maior que o tempo de ocorrência do evento de interesse.

- Censura intervalar: ocorre quando se sabe apenas que o tempo de ocorrência do evento de interesse está em um certo intervalo de tempo.

Segundo Kalbfleisch & Prentice (2002) e Collett (2015), entre outros, o tipo de censura à direita é o que ocorre com maior frequência em problemas práticos. No entanto, situações em que a censura é intervalar também ocorrem com bastante frequência, especialmente na área médica. Peto (1973), por exemplo, relata dados de consultas anuais com 196 meninas, com o objetivo de avaliar o tempo até o desenvolvimento da maturidade sexual, registrado

no momento de cada consulta. Outro exemplo é o tratamento de mulheres com câncer de mama, em que os retornos ao especialista ocorrem a cada quatro a seis meses e, durante as visitas, os médicos avaliam a aparência da paciente, principalmente a retração da mama, a qual apresenta um impacto negativo na aparência e estética geral da paciente, representando o evento de interesse a ser estudado (Finkelstein, 1986). Nos artigos de Kongerud & Samuelsen (1991) e Samuelsen & Kongerud (1993) são relatados dois estudos sobre sintomas respiratórios e asmáticos, respectivamente, de funcionários noruegueses de uma determinada fábrica que trabalhavam com alumínio. Em tais estudos, sabe-se somente que o tempo até o desenvolvimento para ambos sintomas ocorre quando é atestado o diagnóstico após visitas médicas. No estudo sobre a exposição à tuberculose realizado por Smith et al. (1997), os tempos de infecção por tuberculose não são observados de forma exata, havendo a necessidade de se registrar o intervalo de tempo definido pela conversão tuberculínica, caracterizando-se assim um cenário em que a censura é intervalar. Outro exemplo pode ser encontrado em Sun (2006), que apresenta um estudo envolvendo 16 centros, avaliando os tempos de infecção por HIV em hemofílicos, divididos em grupos tratamento e placebo, que são avaliados trimestralmente. Mais exemplos podem ser encontrados em Klein & Moechsberger (2003), Gómez et al. (2004), Sun (2006) e Colosimo & Giolo (2006), entre outros.

Existe uma vasta literatura relacionada aos modelos de sobrevivência para dados com censura intervalar. O artigo de Finkelstein (1986) foi um dos primeiros trabalhos publicados para dados com censura intervalar utilizando o modelo de riscos proporcionais de Cox (1972). Sun (1996) apresentou uma extensão do teste *log rank*, ilustrando com dados de pacientes com AIDS. Mais tarde, Sun (1997) apresenta uma proposta de modelagem para dados com censura intervalar utilizando regressão logística. Artigos de autores como Huang (1996) e Kim (2003) apresentam resultados assintóticos para o modelo de riscos proporcionais, para dados truncados à esquerda. Zhao et al. (2005) apresentam métodos de equação de estimação como procedimento para a obtenção das estimativas dos parâmetros do modelo de riscos proporcionais proposto por Cox (1972), quando tanto os tempos de sobrevivência quanto as covariáveis estão sujeitos a censura intervalar. Em Henschel et al. (2009) é apresentada uma nova proposta de extensão do modelo de riscos proporcionais para acomodar dados de sobrevivência com censura intervalar em que a função de risco basal é modelada via *splines* cúbicos. Entre os artigos de revisão, os trabalhos de Zhang & Sun (2010) e Klein et al. (2013) capítulo 18, apresentam conceitos e métodos sobre censura intervalar e recentes avanços na área.

Um modelo semiparamétrico bastante popular em análise de sobrevivência é o modelo exponencial por partes (MEP). Este modelo, apesar de ser paramétrico em um senso estrito, não requer suposições quanto à forma da função risco, o que é considerado uma vantagem quando comparado a um modelo paramétrico que restringe a forma da função risco. Segundo Ibrahim et al. (2001), o MEP é caracterizado pela aproximação da função risco por segmentos de retas constantes, cujos comprimentos são determinados por uma grade de pontos τ que divide o eixo dos tempos em um número finito de intervalos. A grade τ desempenha um

papel fundamental nas estimativas dos parâmetros de interesse, bem como na qualidade do ajuste do modelo, uma vez que τ determina o número de intervalos para aproximação da função risco.

A literatura relacionada ao MEP é bastante extensa. Friedman (1982) utiliza o MEP para modelar a função risco de base do modelo de Cox (1972), apresentando as condições de existência e a distribuição assintótica dos estimadores de máxima verossimilhança para as taxas de falha e coeficientes de regressão. Em Kim & Proschan (1991), são apresentadas algumas vantagens do MEP, na ausência de covariáveis, comparativamente ao estimador limite produto da função de sobrevivência proposto por Kaplan & Meier (1958). No artigo de Chen & Ibrahim (2001) é empregado o método da máxima verossimilhança para ajustar as taxas de falha do MEP na presença de dados faltantes e fração de curados.

Sob uma perspectiva Bayesiana, Gamerman (1991) estende o modelo de Cox propondo uma abordagem dinâmica para dados de sobrevivência em que os efeitos das covariáveis mudam com o tempo. Posteriormente, Gamerman (1994) apresenta uma abordagem dinâmica (na ausência de covariáveis), ao introduzir uma estrutura de correlação entre as taxas de falha do MEP, associadas aos sucessivos intervalos, através dos hiperparâmetros da distribuição *a priori* dessas taxas. No contexto de modelos de fragilidade, Sahu et al. (1997) utiliza um processo correlacionado *a priori* para as taxas da falha do MEP para modelagem de dados de sobrevivência. Mais detalhes sobre o enfoque Bayesiano para o ajuste de modelos de sobrevivência podem ser encontrados em Ibrahim et al. (2001).

Embora a literatura relacionada ao MEP seja bastante extensa, na maioria dos trabalhos existentes a grade é escolhida de forma arbitrária. Na prática, o problema de especificarmos uma grade apropriada para o ajuste do MEP pode ser resolvido assumindo-se que τ é mais um parâmetro a ser estimado. Observe que sob um enfoque Bsayesiano a extensão é direta, uma vez que basta incluir uma distribuição *a priori* para τ. O primeiro esforço efetivo nesta direção é devido a Arjas & Gasbarra (1994). Estes autores assumem que os pontos finais dos intervalos são definidos de acordo com um processo de saltos baseados na teoria de martingais, incluído no modelo através de distribuições *a priori*. Extensões desta abordagem podem ser encontradas em Mckeague & Tighiouart (2000) e Kim et al. (2006).

Demarqui et al. (2008) apresentam uma abordagem, também Bayesiana, que utiliza a estrutura de agrupamento do modelo partição produto (MPP), proposto por Barry & Hartigan (1992), para modelar diretamente a aleatoriedade de τ. Sob tal abordagem, a modelagem da grade é feita levando-se em conta a disposição dos tempos de falha no eixo do tempo, e a existência de pelo menos um tempo de falha em cada intervalo induzido pela grade aleatória do MEP é garantida. Além disto, apesar de o número de parâmetros a serem estimados poder variar, esta forma de modelagem mantém fixado o número máximo de parâmetros do modelo. Mais recentemente, Demarqui (2010) apresentam extensões do modelo dinâmico proposto em Gamerman (1994) e Gamerman (1991), respectivamente, em que τ é considerado como uma quantidade aleatória. Em seguida, em Demarqui et al. (2014) é proposta uma versão do MEP com grade aleatória para um cenário com presença de fração de curados,

em que neste caso, a especificação da distribuição *a priori* para as taxas de falha do MEP é feita assumindo uma estrutura hierárquica.

Apesar da popularidade do MEP, a literatura relacionada a esse modelo para o ajuste de dados de sobrevivência sujeitos a censura intervalar ainda é relativamente pequena, sendo menor ainda, quando pensamos em tratar τ como mais um parâmetro a ser estimado. Um trabalho bastante citado é o artigo de Lindsey & Ryan (1998), que apresentam um tutorial envolvendo modelos e métodos paramétricos e não paramétricos, em que o MEP é um dos modelos utilizados na modelagem dos dados de sobrevivência com censura intervalar. No artigo de Sinha et al. (1999) é apresentada uma abordagem Bayesiana semiparamétrica assumindo o MEP para modelagem da função de risco basal, e com efeito das covariáveis variando no tempo. O modelo de Sinha et al. (1999) é baseado na ampliação de dados, em que pseudotempos de falha são gerados através de uma distribuição condicional baseada nos próprios intervalos observados e nos valores dos parâmetros de interesse. Sob uma abordagem clássica, Seaman & Bird (2001) utilizaram o MEP para modelar o risco de infecção por HIV pelo uso de drogas injetáveis dentro da prisão com covariáveis mudando no tempo.

Recentemente, Wang et al. (2013) apresentam uma nova proposta de modelagem Bayesiana semiparamétrica, baseada no MEP, para modelar dados de sobrevivência sujeitos a censura intervalar que considera uma estrutura dinâmica entre os coeficientes de regressão, permitindo dessa maneira captar mudanças nos efeitos das covariáveis ao longo do tempo. Assim como em Sinha et al. (1999), o algoritmo de ampliação de dados também é utilizado em Wang et al. (2013), que imputa pseudotempos de falha para cada intervalo de tempo observado. Além disso, a grade τ é tratada como uma quantidade aleatória, e o algoritmo de Monte Carlo via cadeias de Markov com saltos reversíveis (RJMCMC), proposto por Green (1995), é utilizado para amostrar da distribuição *a posteriori*. Independente do tipo de censura, a principal diferença entre as abordagens propostas em Demarqui et al. (2008) e Wang et al. (2013), respectivamente, está relacionada com o algoritmo utilizado para modelar a grade τ, pois em Demarqui et al. (2008) é utilizado a estrutura de agrupamento do MPP ao invés do RJMCMC que é utilizado para amostrar da distribuição *a posteriori*.

Situações em que há a presença de uma fração de curados na população também são recorrentes em dados de sobrevivência sujeitos à censura intervalar. No entanto, a quantidade de trabalhos envolvendo tal cenário é bem menor quando comparado a da censura à direita. Sob uma abordagem frequentista, citamos alguns trabalhos nesta linha, como por exemplo Liu & Shen (2009), que apresentam um modelo de tempos de promoção em que a estimação dos parâmetros de interesse é baseada em uma variação do algoritmo EM para dados com censura intervalar. Posteriormente, Hu & Xiang (2013), propõem um modelo de tempos de promoção semiparamétrico via método de estimação *Spline-Based Sieve*. Lam et al. (2013) consideram um modelo de fragilidade com distribuição contínua positiva e massa no ponto zero para contemplar indivíduos não suscetíveis, caracterizando também a heterogeneidade das condições de saúde dos pacientes suscetíveis. Mais recentemente, Hashimoto et al. (2015), desenvolveram o modelo binomial negativo-Weibull como uma nova forma de modelar dados

de sobrevivência com censura intervalar, baseado no modelo de de Castro et al. (2009). Em um cenário Bayesiano, Banerjee & Carlin (2004) apresentam uma proposta de modelagem completamente paramétrica, baseada no modelo de mistura padrão (Berkson & Gage, 1952), em que é assumida a distribuição de Weibull para modelar os indivíduos não curados, permitindo uma correlação espacial entre os tempos. Thompson & Chhikara (2003) apresentam um modelo de fragilidade Bayesiano com fração de cura baseado no modelo de tempos de promoção em que a distribuição log-normal é utilizada para modelar a distribuição dos tempos de promoção. Sobre o uso do MEP em dados com indivíduos sujeitos a censura intervalar em uma população com a presença de fração de curados, sob uma abordagem frequentista Kim & Jhun (2008) propõem um modelo de fragilidade semiparamétrico baseado no modelo de mistura padrão.

1.1 Objetivos

Esta tese tem como objetivo propor modelos Bayesianos semiparamétricos para dados de sobrevivência sujeitos a censura intervalar. A proposta básica está em desenvolver modelos dinâmicos, baseando-se no MEP com grade aleatória, utilizando a estrutura de agrupamento do MPP, para dados de sobrevivência com censura intervalar, sem e com fração de curados na população em estudo.

Para uma população sem fração de curados, adaptamos o modelo dinâmico proposto por Demarqui et al. (2012) para acomodar dados de sobrevivência com censura intervalar com o efeito das covariáveis variando no tempo. Em situações que envolvem uma fração de curados na população, propomos um novo modelo dinâmico com fração de cura utilizando a abordagem de tempos de promoção (Yakovlev et al., 1993), baseado na extensão do modelo com fração de cura proposto em Demarqui et al. (2014).

1.2 Justificativa

Uma das vantagens de se trabalhar com o MEP é que esse modelo apresenta a flexibilidade dos modelos não paramétricos sem perder as vantagens associadas aos modelos paramétricos, ao não impor restrições sobre a função risco. Artigos como o de Arjas & Gasbarra (1994), Mckeague & Tighiouart (2000), Kim et al. (2006), Demarqui et al. (2008) e Wang et al. (2013) apontam que o MEP com grade aleatória, em geral, fornece melhores resultados que o MEP ajustado através de escolhas *ad-hoc* para a grade τ. Além disso, o uso do MEP com grade aleatória foi pouco explorado em situações para dados de sobrevivência sujeitos a censura intervalar. Desta forma, a generalização da abordagem proposta por Demarqui et al. (2008) para acomodar dados de sobrevivência com censura intervalar tem uma contribuição importante, dada a escassez de trabalhos relacionados ao assunto. Concomitantemente, os artigos de Demarqui et al. (2012) e Demarqui et al. (2014) nos indicam que utilizar uma

estrutura que correlacione as taxas do MEP pode resultar em uma melhora na qualidade de ajuste do modelo aos dados. Portanto, espera-se que os modelos que estão sendo propostos nesta tese possam propiciar melhores resultados para o analista de dados de sobrevivência.

1.3 Organização da tese

Esta tese está organizada da seguinte forma. No Capítulo 2 apresentamos os conceitos básicos em análise de sobrevivência, envolvendo terminologia em análise de sobrevivência, modelo de riscos proporcionais e o modelo com fração de cura. Apresentamos também uma seção específica para dados sujeitos a censura intervalar, onde são discutidos conceitos básicos, função de verossimilhança, para dados com e sem fração de cura, e finalizamos o capítulo com a descrição do algoritmo de ampliação de dados para a imputação dos pseudotempos de falha para os intervalos observados.

No Capítulo 3, apresentamos uma descrição genérica do MPP e em seguida construção do modelo exponencial por partes via estrutura de agrupamento MPP. No Capítulo 4 descrevemos os modelos dinâmicos que discutiremos nesta tese. Para modelos sem fração de curados, apresentamos uma breve descrição das opções disponíveis na literatura, para fins de comparação. Em seguida, é apresentada uma extensão do modelo proposto por Demarqui et al. (2012), capaz de acomodar dados de sobrevivência sujeitos à censura intervalar, apresentando em seguida o algoritmo de ampliação de dados. Na seção seguinte, apresentamos uma extensão do modelo proposto por Demarqui (2010), propondo um novo modelo de fração de cura dinâmico, baseando-o no modelo de tempos de promoção para dados de sobrevivência com censura intervalar.

Apresentamos no Capítulo 5 os resultados dos ajustes dos modelos propostos e os utilizados para a comparação. Neste capítulo dispomos aplicações a dados simulados e reais para modelos sem e com fração de cura, respectivamente. Para os modelos sem fração de cura, utilizamos inicialmente um conjunto de dados simulados e em seguida apresentamos os resultados dos mesmos modelos referente a aplicação aos dados de pacientes com câncer de mama. Por outro, os modelos com fração de cura foram também aplicados a dados simulados e aos dados de soroconversão para o HIV-1 em pacientes hemofílicos. Finalmente, no Capítulo 6 apresentamos as conclusões referentes aos resultados e pesquisas futuras acerca do trabalho desenvolvido nesta tese.

Capítulo 2

Conceitos Gerais

Nas seções seguintes são descritos alguns conceitos importantes que servirão como alicerce para a construção dos modelos desta tese. Na Seção 2.1 apresentamos alguns conceitos básicos sobre análise de sobrevivência. Em seguida, o modelo de riscos proporcionais e o modelo de fração de cura baseado no modelo de tempo de promoção (Yakovlev et al., 1993) são discutidos. Finalmente, apresentamos uma seção relacionada a dados de sobrevivência sujeitos a censura intervalar, descrevendo os conceitos básicos para a construção de modelos e o algoritmo de ampliação de dados para a geração dos pseudotempos de falha de uma maneira geral.

2.1 Conceitos básicos em análise de sobrevivência

Seja T uma variável aleatória contínua e não negativa, correspondendo ao tempo até a ocorrência de um evento de interesse, com função densidade e função de distribuição acumulada denotadas, respectivamente, por $f(t)$ e $F(t)$. Uma das funções mais importantes em análise de sobrevivência, que fornece a probabilidade de um indivíduo sobreviver por mais que um tempo t, é chamada de função de sobrevivência, e é definida por

$$S(t) = 1 - F(t) = P(T > t) = \int_t^\infty f(u)du. \tag{2.1}$$

A função de sobrevivência possui as seguintes propriedades:

1. é monótona não crescente;
2. $S(0) = 1$, isto é, no início do estudo todos os indivíduos ainda estão sob risco;
3. $\lim_{t\to\infty} S(t) = 0$, ou seja, quando o tempo tende ao infinito a probabilidade de falha tende a 1.

Outra função de grande importância em análise de sobrevivência é a função risco, que re-

presenta a taxa de falha instantânea no tempo t, e é definida da seguinte forma

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t}. \tag{2.2}$$

As funções $f(t), S(t), F(t)$ e $h(t)$, associadas à variável aleatória T são matematicamente relacionadas, ou seja, para sabermos a distribuição de T basta conhecermos a expressão de pelo menos uma delas. Dessa forma, usando a função em (2.2), tem-se

$$h(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -\frac{d}{dt}\log(S(t)), \tag{2.3}$$

e a função de sobrevivência em (2.1) pode ser expressa como

$$S(t) = \exp\left(-\int_0^t h(u)du\right) = \exp\left(-H(t)\right), \tag{2.4}$$

em que $H(t)$ é a função risco acumulado. Assim, combinando as relações em (2.3) e (2.4), temos que a função densidade pode ser expressa em termos de $h(t)$ e $S(t)$ da seguinte forma:

$$f(t) = -\frac{d}{dt}S(t) = h(t)S(t) = h(t)\exp\left(-H(t)\right). \tag{2.5}$$

Agora, considerando uma amostra aleatória de n indivíduos, referente a tempos de falha e de censura, denote por T_i como a variável aleatória associada ao tempo de falha e C_i como a variável aleatória associada ao tempo até a ocorrência de censura para o $i-$ésimo indivíduo.

Sob um cenário de censura à direita, $Y_i = \min(T_i, C_i)$ representa o tempo observável e $\delta_i = 1_{\{T_i \leq C_i\}}$ é a variável indicadora de falha ou censura, para $i = 1, 2, \ldots, n$. As funções densidade e sobrevivência para T_i, serão denotadas por $f(\cdot|\boldsymbol{\Theta})$ e $S(\cdot|\boldsymbol{\Theta})$, respectivamente, em que $\boldsymbol{\Theta}$ é o vetor de parâmetros de interesse associado à distribuição do tempo de falha T.

É importante ressaltar que um dos principais interesses em análise de sobrevivência é fazer inferências sobre o vetor de parâmetros $\boldsymbol{\Theta}$, ou seja, aqueles parâmetros que estão associados à distribuição do tempo até a falha. Assim, o vetor de parâmetros associado à distribuição dos tempos de censura é considerado de perturbação.

Então, sob a suposição de que o mecanismo de censura é não informativo, ou seja, que T e C são independentes, a função de verossimilhança é dada por

$$L(\boldsymbol{\Theta}|D) \propto \prod_{i=1}^{n} f(y_i|\boldsymbol{\Theta})^{\delta_i} S(y_i|\boldsymbol{\Theta})^{1-\delta_i},$$

em que $D = \{(y_i, \delta_i) : i = 1, \ldots, n\}$. Utilizando a relação em (2.3), a função de verossimi-

lhança em (2.6) pode ser expressa como

$$L(\boldsymbol{\Theta}|D) \propto \prod_{i=1}^{n} h(y_i|\boldsymbol{\Theta})^{\delta_i} S(y_i|\boldsymbol{\Theta}). \tag{2.6}$$

Em análise de sobrevivência é bastante comum introduzir covariáveis na modelagem para avaliar a relação das mesmas com o risco de ocorrência do evento de interesse. Um dos modelos que utiliza esta estrutura é o modelo de riscos proporcionais proposto por Cox (1972), que será discutido na próxima seção.

2.2 Modelo de riscos proporcionais

A estrutura do modelo de riscos proporcionais possui a seguinte expressão para a função risco

$$h(t|\boldsymbol{\beta}, \boldsymbol{x}) = h_0(t) \exp\left\{\boldsymbol{x}^\top \boldsymbol{\beta}\right\}, \tag{2.7}$$

em que $h_0(t)$ é chamada função risco basal, $\boldsymbol{x}$ é um vetor de dimensão $p \times 1$ de covariáveis e $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ é um vetor $p \times 1$ de coeficientes de regressão.

O modelo de Cox (1972), que utiliza a estrutura do modelo de riscos proporcionais, tem sido amplamente utilizado, principalmente na área médica (Colosimo & Giolo, 2006), devido a flexibilidade que este proporciona ao analista dos dados de sobrevivência. As suposições assumidas acerca deste modelo são de que não há nenhuma restrição sobre a distribuição de probabilidade para os tempos de falha T, consequentemente, não é imposta nenhuma forma analítica conhecida para $h_0(t)$, e dessa forma $h_0(t)$ é modelada não parametricamente. Assume-se também a proporcionalidade dos riscos, ou seja, a razão de funções risco entre os indivíduos i e j,

$$\frac{h(t|\boldsymbol{x}_i)}{h(t|\boldsymbol{x}_j)} = \frac{h_0(t) \exp\left\{\boldsymbol{x}_i^\top \boldsymbol{\beta}\right\}}{h_0(t) \exp\left\{\boldsymbol{x}_j^\top \boldsymbol{\beta}\right\}} = \exp\left\{\boldsymbol{\beta}(\boldsymbol{x}_i^\top - \boldsymbol{x}_j^\top)\right\},$$

não depende do tempo t.

Sem especificar uma forma analítica para a função risco basal, o vetor de coeficientes $\boldsymbol{\beta}$ pode ser estimado utilizando uma função de verossimilhança parcial (Cox, 1972; Kalbfleisch, 1978). Por outro lado, se a função de risco basal $h_0(t)$ for modelada parametricamente, a função de verossimilhança em (2.6) assume a seguinte forma

$$L(\boldsymbol{\Theta}, \boldsymbol{\beta}|D) \propto \prod_{i=1}^{n} \left[h_0(y_i|\boldsymbol{\Theta}) \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right)\right]^{\delta_i} \exp\left\{-H_0(y_i|\boldsymbol{\Theta}) \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right)\right\}, \tag{2.8}$$

em que $D = \{(y_i, \delta_i, \boldsymbol{x}_i) : i = 1, \ldots, n\}$.

Considerando a estrutura do modelo de riscos proporcionais, em que uma forma paramétrica é atribuída para $h_0(t)$, sob a perspectiva bayesiana, inferências sobre os parâmetros de

interesse, $\boldsymbol{\Theta}$ e $\boldsymbol{\beta}$ são realizadas baseando-as na seguinte distribuição *a posteriori* conjunta,

$$p(\boldsymbol{\Theta}, \boldsymbol{\beta} \mid D) \propto \prod_{i=1}^{n} \left[h_0(y_i|\boldsymbol{\Theta}) \exp\left\{\boldsymbol{x}_i^\top \boldsymbol{\beta}\right\}\right]^{\delta_i} \exp\left\{-H_0(y_i|\boldsymbol{\Theta}) \exp\left\{\boldsymbol{x}_i^\top \boldsymbol{\beta}\right\}\right\} p(\boldsymbol{\Theta}, \boldsymbol{\beta}), \qquad (2.9)$$

em que $p(\boldsymbol{\Theta}, \boldsymbol{\beta})$ é a distribuição *a priori* conjunta de $\boldsymbol{\Theta}$ e $\boldsymbol{\beta}$. Como a distribuição *a posteriori* em (2.9) não apresenta forma analítica conhecida, utilizamos como solução ao problema métodos de Monte Carlo via cadeias de Markov (MCMC), mais especificamente o amostrador de Gibbs, que é baseado na geração de amostras da distribuição alvo (2.9), através de suas respectivas distribuições condicionais completas (Geman & Geman, 1984; Gelfand & Smith, 1990).

Independente da distribuição de T, a distribuição condicional completa para o vetor $\boldsymbol{\beta}$ baseada em (2.9), não apresenta uma forma analítica conhecida. Entretanto, pode ser mostrado que a distribuição condicional completa do vetor $\boldsymbol{\beta}$ é log-côncava em cada um de seus componentes (Dellaportas & Smith, 1993). Desta forma, o algoritmo da rejeição adaptativa (ARS), proposto por Gilks & Wild (1992), pode ser utilizado para amostrar o vetor $\boldsymbol{\beta}$.

2.3 Modelo com fração de cura

Algumas vezes, ao analisarmos dados de sobrevivência, percebemos que, após o término do estudo, um número considerável de indivíduos não experimentaram o evento de interesse. Esse fato pode ser uma indicação de que, para uma parcela dos indivíduos sob estudo, o evento de interesse nunca ocorrerá, ou seja, podemos ter uma fração da população que é imune ao evento de interesse (ou que é curada). Esse tipo de situação ocorre em muitos estudos clínicos, como por exemplo, em estudos envolvendo pacientes diagnosticados com diferentes tipos de câncer, nos quais apresentam uma fração de pacientes que são curados, pois tem-se observado um percentual cada vez maior de pacientes considerados curados da doença.

A constatação da existência de uma fração de curados na população pode ser feita de maneira empírica através da inspeção da função de sobrevivência estimada a partir de algum método não paramétrico, como por exemplo o estimador de Kaplan e Meyer (Kaplan & Meier, 1958). Se o gráfico da função de sobrevivência estimada apresentar um platô nos maiores tempos de acompanhamento, temos uma indicação da presença de uma fração de curados na população.

Um modelo simples que incorpora a presença de fração de curados em uma determinada população é o chamado modelo de mistura. Este modelo, originalmente proposto por Boag (1949) e Berkson & Gage (1952), considera uma proporção π de indivíduos imunes ou

curados. Dessa forma, a função de sobrevivência populacional é dada por:

$$S_{pop}(t) = \pi + (1 - \pi)S_{NC}(t), \tag{2.10}$$

em que π é fração de cura e $S_{NC}(t)$ é a função de sobrevivência associada aos elementos não curados na população. Existe um número expressivo de trabalhos envolvendo o modelo de mistura para modelar dados de sobrevivência com fração de cura. Este modelo foi extensivamente discutido, como por exemplo, em Farewell (1982, 1986), Kuk & Chen (1992),Maller (1996) e Sy & Taylor (2001), entre outros.

Segundo Ibrahim et al. (2001), apesar da sua simplicidade e popularidade, o modelo de mistura, em (2.10), apresenta alguns problemas, como por exemplo, na presença de covariáveis:

- não apresenta a propriedade de riscos proporcionais se as covariáveis forem introduzidas por π via função de ligação logística;
- produz distribuições *a posteriori* impróprias para os coeficientes de regressão, mesmos para distribuições conhecidas, como por exemplo, a distribuição uniforme, necessitando fornecer distribuições *a priori* próprias.

O modelo de tempo de promoção é um modelo alternativo ao modelo de mistura que contorna os problemas apresentados anteriormente.

2.3.1 Modelo de tempo de promoção

O modelo de fração de cura considerado neste trabalho de tese é o modelo de tempo de promoção proposto por Yakovlev et al. (1993). O apelo utilizado para a construção deste modelo é totalmente biológico, pois em Yakovlev et al. (1993) estuda-se o tempo de evento para duas características distintas sobre o crescimento do tumor cancerígeno. No entanto, o modelo de tempo de promoção pode ser aplicado para vários tipos de dados de sobrevivência com fração de cura (Ibrahim et al., 2001), como por exemplo, o tempo até a morte de um indivíduo, até a primeira infecção de uma doença, etc. Neste caso, podemos pensar que o modelo é construído baseando-se em M fatores latentes, em que M representa o número de causas ou riscos para a ocorrência de um evento de interesse, caracterizando uma estrutura de riscos competitivos.

Construindo o modelo baseado na motivação biológica, seja M o número de células potencialmente cancerígenas para um indivíduo em uma determinada população. Denote por R_c o tempo até que a c-ésima célula cancerígena produza um câncer detectável. Assumimos que, condicional em M, as variáveis $R_c,\ c = 1, 2, \ldots, M$, são independentes e identicamente distribuídas, com função de distribuição de probabilidade acumulada $F(r)$ e função de sobrevivência $S(r) = 1 - F(r)$. Assumimos também que M é independente de $R_1, R_2, \ldots$ O

tempo até a recidiva é definido como $T = \min(R_0, R_1, R_2, \ldots, R_M)$, em que $P(R_0 = \infty) = 1$. É assumido também que M segue uma distribuição de Poisson com média θ, ou seja,

$$P(M = m) = \frac{e^{-\theta}\theta^m}{m!}, \qquad m = 0, 1, 2, \ldots.$$

Dessa forma, a função de sobrevivência populacional para T é obtida da seguinte forma

$$\begin{aligned} S_{pop}(t) &= P(T > t) = P(\min\{R_0, R_1, \ldots, R_M\} > t, M \geq 0) \\ &= e^{-\theta} + e^{-\theta}\sum_{m=1}^{\infty} \frac{[S(t)\theta]^m}{m!} \\ &= e^{-\theta F(t)}. \end{aligned} \tag{2.11}$$

Note que

$$\lim_{t\to\infty} S_{pop}(t) = \exp(-\theta) = P(M = 0) = \pi,$$

em que π é a fração de cura, ou seja, a probabilidade de um indivíduo apresentar nenhuma célula cancerígena. Assim, como $S_{pop}(t)$ não tende a 0 quando t tende ao infinito, dizemos que $S_{pop}(t)$ é uma função de sobrevivência imprópria (Chen et al., 1999).

Usando as relações em (2.3) e (2.4), temos que a função densidade populacional é dada por

$$\begin{aligned} f_{pop}(t) &= -\frac{d}{dt}\left[S_{pop}(t)\right] = -\frac{d}{dt}\left[e^{-\theta F(t)}\right] \\ &= -\left[e^{-\theta F(t)}\right]\left[-\theta F'(t)\right] = e^{-\theta F(t)}\theta f(t) \\ &= \theta f(t) e^{-\theta F(t)}. \end{aligned} \tag{2.12}$$

É importante observar que $f_{pop}(t)$ também é uma função de densidade imprópria pois é obtida a partir da derivada de $S_{pop}(t)$.

A função risco populacional é dada por

$$h_{pop}(t) = \frac{f_{pop}(t)}{S_{pop}(t)} = \theta f(t). \tag{2.13}$$

Conforme mostrado em Ibrahim et al. (2001), a função de sobrevivência para os não curados, $S_{NC}(t)$ é dada por

$$S_{NC}(t) = P(T > t \mid M \geq 1) = \frac{P(T > t, M \geq 1)}{P(M \geq 1)} = \frac{e^{-\theta F(t)} - e^{-\theta}}{1 - e^{-\theta}}. \tag{2.14}$$

Utilizando as relações descritas na Seção 2.1, temos que a função densidade para os não curados é expressa por

$$f_{NC}(t) = -\frac{d}{dt} S_{NC}(t) = \frac{\theta f(t) e^{-\theta F(t)}}{1 - e^{-\theta}}.$$

Consequentemente, a função risco para os não curados é

$$h_{NC}(t) = \frac{f_{NC}(t)}{S_{NC}(t)} = \frac{\theta f(t) e^{-\theta F(t)}}{e^{-\theta F(t)} - e^{-\theta}}.$$

Existe uma relação biunívoca entre o modelo de tempo de promoção e o modelo de mistura padrão. De fato, como mostrado em Ibrahim et al. (2001), o modelo de tempo de promoção pode ser expresso na forma de um modelo de mistura, em que

$$S_{pop}(t) = \exp(-\theta) + \{1 - \exp(-\theta)\} S_{NC}(t). \tag{2.15}$$

Dessa forma, o modelo de mistura padrão é visto como um caso particular do modelo de tempo de promoção, em que a fração de cura é $\pi = \exp(-\theta)$.

Assim como no modelo de riscos proporcionais, descritos anteriormente, é de interesse para alguns pesquisadores incluir covariáveis na modelagem dos dados a fim de avaliar a relação das mesmas com a fração de cura dos indivíduos. Em geral, as covariáveis são introduzidas pelo parâmetro θ (média de M) utilizando a seguinte relação

$$\theta = \exp(\boldsymbol{z}^\top \boldsymbol{\psi}),$$

em que $\boldsymbol{z}^\top = (1, z_1, \ldots, z_p)^\top$ é um vetor de covariáveis para o i-ésimo indivíduo e $\boldsymbol{\psi} = (\psi_0, \psi_1, \ldots, \psi_p)$ é um vetor $(p+1)$-dimensional de coeficientes de regressão. Dessa forma, a fração de cura é expressa da seguinte forma

$$\pi = \exp\left[-\exp\left(\boldsymbol{z}^\top \boldsymbol{\psi}\right)\right]. \tag{2.16}$$

Reescrevendo a função risco populacional em termos da covariáveis temos que

$$h_{pop}(t \mid \boldsymbol{z}) = \exp(\boldsymbol{z}^\top \boldsymbol{\psi}) f(t). \tag{2.17}$$

Note que a função risco em (2.17) possui a propriedade de riscos proporcionais, pois $h_{pop}(t \mid \boldsymbol{z}_i)/h_{pop}(t \mid \boldsymbol{z}_j)$ não depende do tempo t. De fato, Rodrigues et al. (2009a) mostram que a estrutura de riscos proporcionais é válida somente quando a distribuição da variável latente M é Poisson.

Considere uma amostra aleatória de tamanho n de tempos de falha em que o tipo de censura é a direita. Como já definido, a variável $T_i = \min(R_{i0}, R_{i1}, R_{i2}, \ldots, R_{iM})$ representa o tempo até a recidiva e defina C_i como o tempo até a ocorrência de censura para o $i-$ésimo indivíduo. O tempo observado é definido como $Y_i = \min(T_i, C_i)$ e $\delta_i = I(T_i \leq C_i)$ é a variável indicadora de falha, para $i = 1, 2, \ldots, n$. Assim denote $D = \{(y_i, \delta_i, \boldsymbol{z}_i), i = 1, \ldots, n\}$ como os dados observados e $D_c = \{(y_i, \delta_i, m_i, \boldsymbol{z}_i), i = 1, \ldots, n\}$ como o conjunto dos dados completos, baseados nos dados observados e número de causas latentes.

Sejam $f(\cdot, \boldsymbol{\Theta})$ e $S(\cdot, \boldsymbol{\Theta})$ as funções densidade e de sobrevivência, respectivamente, associ-

adas à R_c, em que $\boldsymbol{\Theta}$ é um vetor de parâmetros da distribuição de probabilidade dos tempo de promoção R_c. Assumindo que o mecanismo de censura é não informativo, temos que a função de verossimilhança para dos dados completos é dada por

$$\begin{aligned} L(\boldsymbol{\Theta}, \boldsymbol{\psi} \mid D_c) &= \prod_{i=1}^{n} \left[m_i h(y_i \mid \boldsymbol{\Theta})\right]^{\delta_i} S(y_i \mid \boldsymbol{\Theta})^{m_i} \\ &\quad \times \exp\left\{\sum_{i=1}^{n} m_i \boldsymbol{z}_i^\top \boldsymbol{\psi} - \log(m_i!) - e^{\boldsymbol{z}_i^\top \boldsymbol{\psi}}\right\}. \end{aligned} \tag{2.18}$$

A função de verossimilhança para os dados observados é obtida somando nos valores de M. Assim,

$$\begin{aligned} L(\boldsymbol{\Theta}, \boldsymbol{\psi} \mid D) &= \sum_{\mathbf{M}} \prod_{i=1}^{n} \left\{\left[m_i f(y_i \mid \boldsymbol{\Theta})\right]^{\delta_i} S(y_i \mid \boldsymbol{\Theta})^{m_i - \delta_i}\right\} \frac{e^{-e^{\boldsymbol{z}_i^\top \boldsymbol{\psi}}} e^{\boldsymbol{z}_i^\top \boldsymbol{\psi} m_i}}{m_i!} \\ &= \prod_{i=1}^{n} f_{pop}(y_i \mid \boldsymbol{\Theta}, \boldsymbol{\psi}, \boldsymbol{z}_i)^{\delta_i} S_{pop}(y_i \mid \boldsymbol{\Theta}, \boldsymbol{\psi}, \boldsymbol{z}_i)^{1-\delta_i}. \end{aligned} \tag{2.19}$$

Substituindo (2.11) e (2.12) em (2.19), a função de verossimilhança observada é dada por

$$L(\boldsymbol{\Theta}, \boldsymbol{\psi} \mid D) = \prod_{i=1}^{n} \left[\theta_i f(y_i \mid \boldsymbol{\Theta}) e^{-\theta_i F(y_i \mid \boldsymbol{\Theta})}\right]^{\delta_i} \left[e^{-\theta_i F(y_i \mid \boldsymbol{\Theta})}\right]^{1-\delta_i}. \tag{2.20}$$

Combinando a função de verossimilhança (2.19) com a distribuição *a priori* conjunta $p(\boldsymbol{\Theta}, \boldsymbol{\psi})$, temos que a distribuição *a posteriori* conjunta para o modelo de tempo de promoção tem a forma,

$$\begin{aligned} p(\boldsymbol{\Theta}, \boldsymbol{\psi} \mid D) &\propto \prod_{i=1}^{n} f_{pop}(y_i \mid \boldsymbol{\Theta}, \boldsymbol{\psi})^{\delta_i} S_{pop}(y_i \mid \boldsymbol{\Theta}, \boldsymbol{\psi})^{1-\delta_i} p(\boldsymbol{\Theta}, \boldsymbol{\psi}) \\ &\propto \prod_{i=1}^{n} \left[\theta_i f(y_i \mid \boldsymbol{\Theta}) e^{-\theta_i F(y_i \mid \boldsymbol{\Theta})}\right]^{\delta_i} \left[e^{-\theta_i F(y_i \mid \boldsymbol{\Theta})}\right]^{1-\delta_i} p(\boldsymbol{\Theta}, \boldsymbol{\psi}). \end{aligned} \tag{2.21}$$

Obter as estimativas *a posteriori* para $\boldsymbol{\Theta}$ e $\boldsymbol{\psi}$ através de (2.21) não é uma tarefa fácil. Então, baseamos a distribuição *a posteriori* no vetor de variáveis latentes $\boldsymbol{M}$ para obter expressões mais simples do ponto de vista do tratamento analítico (Ibrahim et al., 2001) e obtemos a seguinte expressão:

$$\begin{aligned} p(\boldsymbol{\Theta}, \boldsymbol{\psi}, \mathbf{M} \mid D) &\propto \prod_{i=1}^{n} \left[m_i f(y_i \mid \boldsymbol{\Theta})\right]^{\delta_i} S(y_i \mid \boldsymbol{\Theta})^{m_i - \delta_i} \\ &\quad \times \exp\left(\sum_{i=1}^{n} m_i \boldsymbol{z}_i^\top \boldsymbol{\psi} - \log(m_i!) - e^{\boldsymbol{z}_i^\top \boldsymbol{\psi}}\right) p(\boldsymbol{\Theta}, \boldsymbol{\psi}). \end{aligned} \tag{2.22}$$

Dessa forma, dividiremos a amostragem de $(\boldsymbol{\Theta}, \boldsymbol{\psi}, \mathbf{M} \mid D)$, em duas distribuições condicionais, $(\boldsymbol{\Theta}, \mathbf{M} \mid \boldsymbol{\psi}, D)$ e $(\boldsymbol{\psi} \mid \boldsymbol{\Theta}, D)$ obtidas a partir da expressão (2.21). Tal estratégia representa o amostrador de Gibbs colapsado proposto por Liu (1994), pois a distribuição condicional completa para o vetor $\boldsymbol{\psi}$ é obtida utilizando a função de verossimilhança somada no vetor do número de causas $\boldsymbol{M}$.

Como visto em Ibrahim et al. (2001), independente da distribuição de probabilidade assumida para o tempo de promoção, pode ser mostrado que a distribuição condicional completa para o número de causas latentes é dada por

$$[M_i \mid \boldsymbol{\Theta}, \boldsymbol{\psi}, D] \sim \text{Poisson}\left(S(y_i \mid \boldsymbol{\Theta}) \exp(\boldsymbol{z}_i^\top \boldsymbol{\psi})\right) + \delta_i, \ i = 1, \ldots, n.$$

Neste caso, uma importante informação é que a propriedade de log-concavidade é garantida para cada componente da distribuição condicional completa de $\boldsymbol{\psi}$ (Dellaportas & Smith, 1993), dessa forma utilizamos o ARS para gerar da distribuição condicional completa de $\boldsymbol{\psi}$.

2.4 Censura Intervalar

2.4.1 Conceitos iniciais

Para dados sujeitos a censura intervalar, tem-se somente o conhecimento de que os tempos de falha T não são mais observados de forma exata, ou seja, sabe-se somente que o evento de interesse ocorreu em algum momento de um intervalo de tempo do tipo $(L, R]$, em $L < T \leq R$. Quando $L = R$, observamos um tempo de falha de forma exata.

Na prática, os valores de L e R referem-se a tempos de avaliação e, assim, para alguns indivíduos o evento poderá ocorrer após a última visita com o avaliador, caracterizando observações sujeitas a censura à direita e dessa forma o tempo de falha T pode ocorrer no intervalo (L, ∞). De maneira similar, para alguns indivíduos o evento de interesse pode ter ocorrido anteriormente a primeira avaliação, representando indivíduos sujeitos a censura à esquerda, ou seja, o intervalo observado é $(0, R]$ com $L = 0$ que representa o início do estudo e R o tempo decorrido desde o início até a primeira visita. Na Tabela 2.1 é apresentado um resumo sobre a estrutura de dados com censura intervalar e seus casos especiais.

Tabela 2.1: Resumo de cenários para dados censurados.

Tempo de falha exato	$T = t$
Censurado à direita	$T \in (L, \infty)$
Censurado à esquerda	$T \in (0, R]$
Com tempo intervalar	$T \in (L, R]$

Sun (2006) apresenta diferentes estruturas dos dados de sobrevivência com censura intervalar e apresentamos a seguir alguns casos que geralmente ocorrem na prática.

A censura intervalar denominada do caso I ocorre quando somente sabemos que T é maior ou menor do que um tempo de monitoramento U. Dessa forma, temos o par $(T_i, U_i, i = 1, \ldots, n)$, assumindo que T_i é independente de U_i. O elemento i é inspecionado no tempo U_i e $\delta_i = 1_{\{T_i \leq U_i\}}$ nos indica o estado atual do elemento com respeito a um determinado evento. Por esse motivo, a censura intervalar do caso I é frequentemente denominada de dados de estado corrente (*current status data*). Sendo assim, o intervalo observado é expresso da seguinte forma:

$$(L_i, R_i] = \begin{cases} (0, U_i], & \text{se } T_i \leq U_i \\ (U_i, \infty), & \text{se } T_i > U_i. \end{cases}$$

Ilustramos o caso I pela Figura 2.1 a seguir. Note que neste caso o indivíduo não foi avaliado após o tempo de monitoramento U, apresentando um intervalo observado $(L = 0; R = u]$.

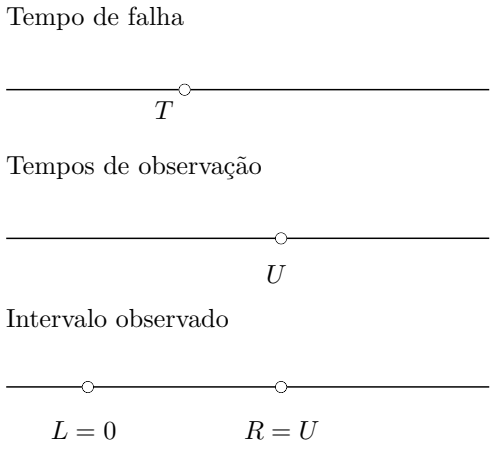

Figura 2.1: Ilustração de dados de sobrevivência com censura intervalar do caso I.

Hoel & Walburg (1972) utilizaram dados sobre o tempo de aparecimento de câncer de pulmão em 144 camundongos machos com uma taxa de câncer muito alta, como um exemplo de censura intervalar do caso I. Os camundongos foram divididos em dois grupos: ambiente convencional (96 ratos) e um ambiente livre de germes (48 ratos). Os tumores pulmonares foram assumidos como sendo não letais. Ao realizar o sacrifício do camundongo, em um tempo aleatório, os pesquisadores avaliavam se o animal apresentava ou não tumores no pulmão. Se os pesquisadores descobrissem tumores pulmonares no momento do sacrifício, era sabido então sobre o aparecimento do câncer de pulmão, que havia ocorrido em algum tempo anterior. Por outro lado, se houvesse a ausência de tumores no momento do sacrifício tinha-se o conhecimento que o aparecimento do câncer poderia ocorrer em algum tempo depois, ou talvez não. Observe que em ambas situações os pesquisadores não observaram o aparecimento de câncer exatamente, caracterizando um cenário de censura intervalar no caso I.

Suponha agora que fixamos dois tempos de observação denotados por U e V com $U < V$, ambos observados para cada indivíduo, representando dados com censura intervalar no caso II. Note que, se $U = V$ regredimos ao caso I. Sabemos somente que o tempo de falha T

ocorre antes do primeiro tempo de monitoramento U, entre os dois tempos de observação ou após o segundo tempo V. O intervalo observado para o i-ésimo indivíduo pode ser escrito da seguinte forma:

$$(L_i, R_i] = \begin{cases} (0, U_i], & \text{se } T_i \leq U_i \\ (U_i, V_i], & \text{se } T_i \in (U_i, V_i] \\ (V_i, \infty), & \text{se } T_i > V_i. \end{cases}$$

A Figura 2.1 mostra um exemplo que ilustra o caso II. Podemos perceber que o tempos de falha pertence aos dois tempos de monitoramento U e V, resultando em um intervalo observado igual a $(L = U; R = V]$.

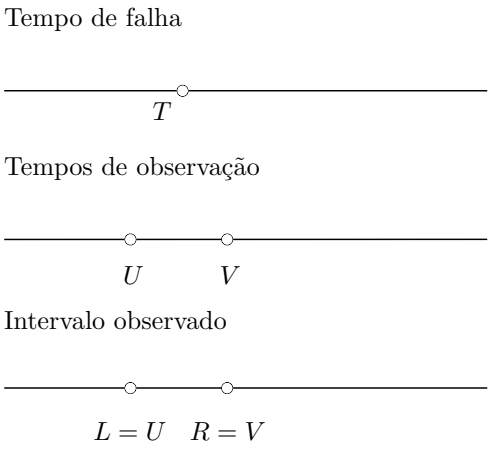

Figura 2.2: Ilustração de dados de sobrevivência com censura intervalar do caso II.

Como ilustração de uma situação em que dados de sobrevivência com censura intervalar são do caso II, suponha que estamos interessados na distribuição dos tempos de infecção por HIV em uma determinada população. Para tal situação, obtemos uma amostra representativa desta população, e que será acompanhada por um período de 2 anos, sendo administrados dois testes para avaliar a infecção HIV em dois tempos distintos do período determinado.

Uma generalização de dados no caso I e II acontece quando existem K tempos de observação fixados por indivíduo, $O_1, O_2, \ldots, O_K$. Neste caso, o evento de interesse ocorre em dois tempos de inspeção consecutivos O_l e O_{l+1}. Dessa forma, observamos um intervalo do tipo $(O_l, O_{l+1}]$. O vetor de tempos observados é determinado por

$$(L_i, R_i] = (0, O_{i1}]1_{\{T_i \leq O_{i1}\}} + \sum_{l=2}^{K} (O_{i(l-1)}, O_{il}]1_{\{O_{i(l-1)} < T_i \leq O_{il}\}} + (O_{iK}, \infty)1_{\{T_i > O_{iK}\}},$$

para $i = 1, 2, \ldots, n$. Gómez et al. (2004) denominam esta situação como caso K.

De maneira geral, ilustramos o caso K pela Figura 2.3. Observe aqui que definimos $K = 6$ tempos de observação para um determinado indivíduo. Vimos que o tempo real de falha T pertence ao intervalo de inspeção O_2 e O_3, dessa forma, temos que o intervalo observado é $(L = O_2, R = O_3]$.

No trabalho de Finkelstein & Wolfe (1985) são apresentados dados envolvendo pacientes

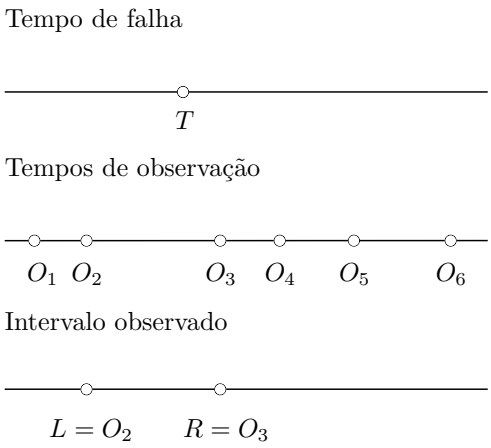

Figura 2.3: Ilustração de dados de sobrevivência com censura intervalar do caso K.

com câncer de mama em que consideraram dois grupos de tratamentos: somente radioterapia ou radioterapia com quimioterapia. As pacientes foram atendidas em várias visitas à clínica do especialista. Nesta situação, o número de visitas clínicas pode variar de paciente para paciente, e em uma das visitas, o médico diagnostica a retração da mama, que é um efeito colateral afetando a aparência da paciente.

Em situações em que os indivíduos são inspecionados nos mesmos tempos de monitoramento, dizemos que os dados são agrupados. Os dados de tempo até a seca de mangueiras é um exemplo de dados agrupados (Colosimo & Giolo, 2006), pois nesta situação, as mangueiras são examinadas no mesmo tempo de observação para serem avaliadas com respeito à presença ou não de praga.

No oitavo capítulo do livro de Sun (2006) é apresentado um outro tipo de censura intervalar chamado censura intervalar dupla. Neste caso, o tempo de interesse T é determinado pela diferença entre dois tempos relacionados, um tempo inicial X e um tempo subsequente B, ou seja,

$$T = B - X$$

em que $X \in (L, R]$ e $B \in (U, V]$. Note que, se $V = \infty$, o tempo B é uma censura a direita, reduzindo ao caso usual de censura intervalar. Segundo Sun (2006), geralmente esta característica para censura intervalar pode aparecer em estudos de progressão de doenças em que dois eventos podem ocorrer, um representando o tempo de infecção e em seguida o tempo do aparecimento de uma determinada doença. Como por exemplo, X representando o tempo de infecção do HIV e B como o tempo até o diagnóstico da AIDS.

2.4.2 Função de verossimilhança e ampliação de dados

De maneira geral, assumindo que o mecanismo gerador da censura é não informativo, a função de verossimilhança para dados com censura intervalar é dada por:

$$L(\boldsymbol{\Theta}, \boldsymbol{\beta} \mid D) = \prod_{i=1}^{n} [S(l_i|\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{x}_i) - S(r_i|\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{x}_i)]^{\delta_i} [S(l_i|\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{x}_i)]^{1-\delta_i}, \tag{2.23}$$

em que $D = \{(l_i, r_i, \delta_i, \boldsymbol{x}_i) : i = 1, 2, \ldots, n\}$ representa o conjunto dos dados observados, $S(\cdot|\boldsymbol{\Theta})$ é a função de sobrevivência e δ_i é o indicador de falha.

Dessa forma, a distribuição *a posteriori* para $\boldsymbol{\Theta}$ e $\boldsymbol{\beta}$ é expressa da seguinte forma

$$\begin{aligned} p(\boldsymbol{\Theta}, \boldsymbol{\beta} \mid D) &\propto L(\boldsymbol{\Theta}, \boldsymbol{\beta} \mid D)\, p(\boldsymbol{\Theta}, \boldsymbol{\beta}) \\ &\propto \left\{ \prod_{i=1}^{n} [S(l_i|\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{x}_i) - S(r_i|\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{x}_i)]^{\delta_i} [S(l_i|\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{x}_i)]^{1-\delta_i} \right\} \\ &\quad \times p(\boldsymbol{\Theta}, \boldsymbol{\beta}), \end{aligned} \tag{2.24}$$

em que $p(\boldsymbol{\Theta}, \boldsymbol{\beta})$ representa a distribuição *a priori* conjunta para $\boldsymbol{\Theta}$ e $\boldsymbol{\beta}$. A função de verossimilhança dada em (2.23) assume uma forma complicada, envolvendo um produto de diferenças de funções de sobrevivência, trazendo dificuldades do ponto de vista da implementação computacional, tanto na abordagem clássica quanto Bayesiana.

Uma estratégia alternativa, e que se mostra bastante atrativa, consiste em imputar os tempos de sobrevivência, o que contorna o problema de utilizar a função de verossimilhança em (2.23) além de propiciar uma distribuição *a posteriori* mais tratável, tanto do ponto de vista analítico quanto do ponto de vista computacional. Sob esta abordagem, modelos desenvolvidos para dados com censura à direita podem ser utilizados num contexto de censura intervalar.

Em Sun (2006), no Capítulo 2, são apresentadas duas formas de se trabalhar dados com censura intervalar, tomando um tempo de sobrevivência exato para cada intervalo observado. Na primeira, sabendo que o evento de interesse ocorreu em um intervalo de tempo, assume-se como tempo observado o limite inferior, o limite superior ou então o ponto médio do intervalo, a fim de viabilizar a aplicação de métodos para dados com censura à direita, uma vez que existem poucos pacotes que acomodam dados com censura intervalar. No entanto, autores como Rücker & Messerer (1988), Odell et al. (1992) e Dorey et al. (1993), ressaltam que tempos imputados para dados com censura intervalar utilizando o método mencionado acima podem conduzir a vícios para a estimação dos parâmetros de interesse. A segunda forma consiste em estimar um tempo de sobrevivência exato através de uma distribuição condicional nos intervalos de tempos observados e nos valores dos parâmetros, que descrevemos com mais detalhes na Seção 2.4.3. Dessa forma, definimos D_c como os dados completos, ou os dados aumentados, em que $D_c = \{(y_i, \delta_i) : i = 1, \ldots, n\}$. Note que y_i é um pseudotempo de falha para cada indivíduo que tenha experimentado o evento de interesse,

isto é, y_i corresponde a um tempo imputado para $(l_i, r_i]$, se $\delta_i = 1$, e $y_i = l_i$ caso contrário, para $i = 1, \ldots, n$.

A utilização do algoritmo de ampliação de dados consiste em obter uma forma simples para $L(\boldsymbol{\Theta} \mid D_c)$ e dessa forma amostrar da distribuição *a posteriori* $p(\boldsymbol{\Theta} \mid D_c)$ e da distribuição condicional $p(T \mid D, \boldsymbol{\Theta})$, tornando um processo muito mais fácil do que amostrar diretamente de (2.24). Uma discussão mais detalhada sobre o uso do algoritmo de ampliação de dados, não somente num contexto de análise de sobrevivência, pode ser encontrada em Tanner & Wong (1987), Tanner (1991), Wei & Tanner (1991) e van Dyk & Meng (2001).

Assim, ao utilizar o algoritmo de ampliação de dados para dados de sobrevivência sem fração de curados, a função de verossimilhança em (2.23) poder ser reexpressa da seguinte forma

$$L(\boldsymbol{\Theta}, \boldsymbol{\beta} | D_c) = \prod_{i=1}^{n} [h(y_i|\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{x}_i)]^{\delta_i} \exp\{-H(y_i|\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{x}_i)\}, \tag{2.25}$$

como $D_c = \{(y_i, \delta_i, \boldsymbol{x}_i) : i = 1, 2, \ldots, n\}$, assumindo a mesma formulação como em (2.8), para um cenário em que a censura é à direita. Assim, a distribuição *a posteriori* para $\boldsymbol{\Theta}$ e $\boldsymbol{\beta}$ segue da mesma forma como descrito na Seção 2.2.

Para um cenário com fração de curados na população, temos que a função de verossimilhança para dados de sobrevivência sujeitos a censura intervalar é expressa da seguinte maneira

$$L(\boldsymbol{\Theta}, \boldsymbol{\psi} \mid D) = \prod_{i=1}^{n} [S_{pop}(l_i|\boldsymbol{\Theta}, \boldsymbol{\psi}, \boldsymbol{z}_i) - S_{pop}(r_i|\boldsymbol{\Theta}, \boldsymbol{\psi}, \boldsymbol{z}_i)]^{\delta_i} [S_{pop}(l_i|\boldsymbol{\Theta}, \boldsymbol{\psi}, \boldsymbol{z}_i)]^{1-\delta_i}, \tag{2.26}$$

em que $D = \{(l_i, r_i, \delta_i, \boldsymbol{z}_i) : i = 1, 2, \ldots, n\}$. A demonstração deste resultado pode ser encontrada em Hashimoto et al. (2015). Assumindo a estrutura do modelo de tempos de promoção descrito na Seção 2.3, a função de verossimilhança em (2.26) é escrita da seguinte forma

$$L(\boldsymbol{\Theta}, \boldsymbol{\psi} \mid D) = \prod_{i=1}^{n} \left[e^{-\exp\left(\boldsymbol{z}_i^\top \boldsymbol{\psi}\right)(1-S(l_i|\boldsymbol{\Theta}))} - e^{-\exp\left(\boldsymbol{z}_i^\top \boldsymbol{\psi}\right)(1-S(r_i|\boldsymbol{\Theta}))} \right]^{\delta i} \left[e^{-\exp\left(\boldsymbol{z}_i^\top \boldsymbol{\psi}\right)(1-S(l_i|\boldsymbol{\Theta}))} \right]^{1-\delta i}. \tag{2.27}$$

Ao utilizar o algoritmo de ampliação de dados, temos que a função de verossimilhança baseada nos dados aumentados assume a forma

$$\begin{aligned} L(\boldsymbol{\Theta}, \boldsymbol{\psi} \mid D_c) &= \prod_{i=1}^{n} f_{pop}(y_i \mid \boldsymbol{\Theta}, \boldsymbol{\psi})^{\delta_i} S_{pop}(y_i \mid \boldsymbol{\Theta}, \boldsymbol{\psi})^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left\{ \theta_i f(y_i \mid \boldsymbol{\Theta}) e^{-\theta_i F(y_i|\boldsymbol{\Theta})} \right\}^{\delta_i} \left\{ e^{-\theta_i F(y_i|\boldsymbol{\Theta})} \right\}^{1-\delta_i}, \end{aligned} \tag{2.28}$$

em que $D_c = \{(y_i, \delta_i, \boldsymbol{z}_i), i = 1, 2, \ldots, n\}$. Note que a função acima é análoga a função de verossimilhança para os dados observados em (2.19), com T sendo uma variável aleatória

observável. Na presença dos pseudotempos de falha, a função de verossimilhança condicional no número de causas M é dada por

$$L(\boldsymbol{\Theta}, \boldsymbol{\psi} \mid D_c) = \prod_{i=1}^{n} [m_i f(y_i \mid \boldsymbol{\Theta})]^{\delta_i} S(y_i \mid \boldsymbol{\Theta})^{m_i - \delta_i} \times \exp\left[\sum_{i=1}^{n} m_i \boldsymbol{z}_i^{\top} \boldsymbol{\psi} - \log(m_i!) - e^{\boldsymbol{z}_i^{\top} \boldsymbol{\psi}}\right], \quad (2.29)$$

em que $D_c^* = \{(y_i, \delta_i, \boldsymbol{z}_i, m_i), i = 1, 2, \ldots, n\}$, em que $y_i \mid \delta_i = 1$ e m_i são valores não observáveis, ou seja, quantidades aleatórias. Note que a verossimilhança acima apresenta o mesmo formato da expressão em (2.18), representando um formato bem mais atrativo quando comparado ao da expressão 2.27.

2.4.3 Algoritmo de ampliação de dados

Nesta seção apresentamos, de maneira geral, o algoritmo de ampliação de dados para a geração dos pseudotempos de falha. Vimos que para dados de sobrevivência sujeitos à censura intervalar os dados observados são definidos como $D = \{(l_i, r_i, \delta_i, x_i) : i = 1, 2, \ldots, n\}$ e T_i como o tempo de falha não observável associado ao $i-$ésimo indivíduo que ocorreram em intervalos do tipo $(l_i, r_i]$, para o $i-$ésimo indivíduo, com $l_i \leq r_i$, para $i = 1, \ldots, n$.

Seguindo Gómez et al. (2004), assumindo que T tem função de distribuição $F(t \mid \boldsymbol{\Theta}, \boldsymbol{\beta})$ e função densidade $f(t \mid \boldsymbol{\Theta}, \boldsymbol{\beta})$, temos que a distribuição condicional nos dados observados $p(T \mid \boldsymbol{\Theta}, \boldsymbol{\beta}, D)$ para os pseudotempos de falha é dada por

$$q(t_i \mid \boldsymbol{\Theta}, \boldsymbol{\beta}, D) = \frac{f(t_i \mid \boldsymbol{\Theta}, \boldsymbol{\beta})}{F(r_i \mid \boldsymbol{\Theta}, \boldsymbol{\beta}) - F(l_i \mid \boldsymbol{\Theta}, \boldsymbol{\beta})},$$

com $l_i \leq t_i \leq r_i, \quad i = 1, \ldots, n$ e $q(t_i \mid \boldsymbol{\Theta}, \boldsymbol{\beta}, D)$ é uma função densidade de probabilidade truncada nos intervalos observados $[l_i, r_i]$. A função de distribuição acumulada truncada em $[l_i, r_i]$ é dada por

$$Q(t_i \mid \boldsymbol{\Theta}, \boldsymbol{\beta}, D) = \begin{cases} \dfrac{F(t_i \mid \boldsymbol{\Theta}, \boldsymbol{\beta}) - F(l_i \mid \boldsymbol{\Theta}, \boldsymbol{\beta})}{F(r_i \mid \boldsymbol{\Theta}, \boldsymbol{\beta}) - F(l_i \mid \boldsymbol{\Theta}, \boldsymbol{\beta})}, & l_i \leq t_i \leq r_i, \\ 0, & \text{caso contrário.} \end{cases}$$

$i = 1, \ldots, n$. Dessa forma, pelo método da transformação inversa temos que

$$v_i = \frac{F(t_i \mid \boldsymbol{\Theta}, \boldsymbol{\beta}) - F(l_i \mid \boldsymbol{\Theta}, \boldsymbol{\beta})}{F(r_i \mid \boldsymbol{\Theta}, \boldsymbol{\beta}) - F(l_i \mid \boldsymbol{\Theta}, \boldsymbol{\beta})} \quad \Leftrightarrow \quad F(t_i \mid \boldsymbol{\Theta}, \boldsymbol{\beta}) = \underbrace{F(l_i \mid \boldsymbol{\Theta}, \boldsymbol{\beta}) + v_i [F(r_i \mid \boldsymbol{\Theta}, \boldsymbol{\beta}) - F(l_i \mid \boldsymbol{\Theta}, \boldsymbol{\beta})]}_{h_i},$$

em que $v_i \sim Unif[0, 1]$. Assim, se F tenha inversa, os pseudotempos de falha são gerados da seguinte forma

$$t_i = F^{-1}(h_i \mid \boldsymbol{\Theta}, \boldsymbol{\beta}, D), i = 1, \ldots, n.$$

Uma estratégia equivalente consiste gerar os pseudotempos de falha utilizando a seguinte expressão:

$$t_i = F^{-1}(w_i \mid \boldsymbol{\Theta}, \boldsymbol{\beta}, D), i = 1, \ldots, n,$$

em que

$$w_i \sim Unif\left[F(l_i \mid \boldsymbol{\Theta}, \boldsymbol{\beta}), F(r_i \mid \boldsymbol{\Theta}, \boldsymbol{\beta})\right].$$

Dessa forma, o amostrador de Gibbs tem os seguintes passos:

1. Gerar $w_i \sim Unif\left[F(l_i \mid \boldsymbol{\Theta}, \boldsymbol{\beta}), F(r_i \mid \boldsymbol{\Theta}, \boldsymbol{\beta})\right]$ para intervalos observados com $\delta_i = 1$;
2. Atualizar o vetor de parâmetros de interesse $\boldsymbol{\Theta}$ e $\boldsymbol{\beta}$, condicional nos pesudotempos de falha e nos dados observados.
3. Retornar ao passo 1 até a convergência da cadeia.

Uma vez gerados os pseudotempos de falha temos que os dados completos (ou ampliados), são da seguinte forma:

$$y_i = \begin{cases} t_i, & \text{se } \delta_i = 1 \\ l_i, & \text{se } \delta_i = 0 \end{cases}$$

sendo t_i e l_i o $i-$ésimo pseudo-tempo de falha gerado e o limite inferior do intervalo observado, respectivamente.

Capítulo 3

Modelo exponencial por partes

Segundo Ibrahim et al. (2001), o MEP é um dos modelos mais populares na modelagem semiparamétrica de dados de sobrevivência, pois apesar de ser paramétrico em um senso estrito, não impõe restrições quanto a forma da função risco, diferentemente de outros modelos paramétricos como: exponencial, Weibull e lognormal, entre outros.

Como visto em Ibrahim et al. (2001) o MEP é construído com base em uma aproximação da função risco por segmentos de retas, cujos comprimentos são determinados por uma grade de tempos τ que divide o eixo dos tempos em um número finito de intervalos. Matematicamente, a grade de tempos é definida como $\tau = \{s_0, s_1, ..., s_b\}$, em que $0 = s_0 < s_1 < ... < s_b = \infty$, que induz intervalos da forma $I_j = (s_{j-1}, s_j]$, para $j = 1, 2, ..., b$.

A função risco é definida da seguinte forma:

$$h(t|\boldsymbol{\lambda}, \tau) = \lambda_j, \quad \text{se} \quad t \in I_j, \quad \lambda_j > 0 \quad j = 1, ..., b, \tag{3.1}$$

em que $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_b)^\top$.

Segundo Demarqui (2010), para a obtenção de $H(t)$ e $S(t)$, é conveniente definir

$$t_j = \begin{cases} s_{j-1}, & \text{se} \quad t \leq s_{j-1}, \\ t, & \text{se} \quad t \in I_j, \\ s_j, & \text{se} \quad t > s_j, \end{cases} \tag{3.2}$$

para $j = 1, \ldots, b$. Então, utilizando as equações (3.1) e (3.2), a função risco acumulado pode ser escrito da seguinte forma

$$H(t|\boldsymbol{\lambda}, \tau) = \sum_{j=1}^{b} \lambda_j (t_j - s_{j-1}). \tag{3.3}$$

Logo, a função de sobrevivência pode ser escrita como

$$S(t|\boldsymbol{\lambda}, \tau) = \exp\left\{ - \sum_{j=1}^{b} \lambda_j (t_j - s_{j-1}) \right\}. \tag{3.4}$$

Assim, dizemos que T segue uma distribuição exponencial por partes com grade de tempos τ e vetor de taxas de falha $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_b)^\top$, denotada por $T \sim MEP(\boldsymbol{\lambda}, \tau)$, se sua função densidade é dada por

$$f(t|\boldsymbol{\lambda}, \tau) = \lambda_j \exp\left\{-\sum_{j=1}^{b} \lambda_j (t_j - s_{j-1})\right\},$$

para $t \in I_j$, $j = 1, 2, ..., b$. Observe que se $b = 1$, o modelo se reduz ao modelo exponencial.

Apesar da grande flexibilidade do MEP, o impacto da escolha de τ tem um papel muito importante na estimação dos parâmetros de interesse, bem como na qualidade do ajuste do modelo. Esta questão tem sido um dos grandes desafios para se trabalhar com o MEP. Embora exista uma vasta literatura relacionada a este modelo, na maioria dos trabalhos a escolha de τ é feita de maneira arbitrária. Algumas discussões sobre a escolha da grade podem ser encontradas em Kalbfleisch (1973), Breslow (1974) e Sahu et al. (1997), entre outros.

Uma solução para evitarmos de escolher arbitrariamente τ é encontrada em Demarqui et al. (2008), que propuseram uma extensão do MEP, utilizando a estrutura de agrupamento do Modelo Partição Produto para modelar a aleatoriedade de τ, quando os dados de sobrevivência são sujeitos à censura a direita. A seguir apresentaremos de maneira resumida a formulação geral do modelo MPP e posteriormente mostraremos como sua estrutura de agrupamento pode ser utilizada para modelar o MEP com grade aleatória.

3.1 Modelo partição produto

A seguir apresentamos uma breve revisão sobre o Modelo Partição Produto, proposto por Barry & Hartigan (1992), para identificar pontos de mudança em uma sequência de observações obtida em pontos consecutivos no tempo. Uma discussão mais detalhada sobre este modelo pode ser encontrada em Barry & Hartigan (1992), e Loschi & Cruz (2005), e Quintana & Iglesias (2003), entre outros.

Considere $\mathbf{T} = (T_1, T_2, \ldots, T_n)$ um vetor de variáveis aleatórias obtidas sequencialmente no tempo, e denote por $\mathcal{I} = \{1, 2, \ldots, n\}$ o conjunto de índices associados a $\mathbf{T}$. Seja $\rho = \{i_0, i_1, ..., i_b\}$ uma partição aleatória relacionada ao conjuntos dos índices $\mathcal{I}$, satisfazendo $0 = i_0 < i_1 < ... < i_b = n$, de tal forma que cada partição aleatória ρ divida $\mathbf{T}$ em subsequências (blocos) contíguas, denotadas por

$$\mathbf{T}_{[i_{j-1}, i_j]} = \left(T_{i_{j-1}+1}, \ldots, T_{i_j}\right),$$

para $j = 1, \ldots, b$. Para facilitar o entendimento utilizaremos o exemplo descrito em Demarqui (2008) sobre possíveis relações da partição aleatória ρ com o conjunto de observações $\mathbf{T}$. Na Tabela 3.1 apresentamos as relações de uma partição aleatória ρ com uma sequência $\mathbf{T} = (T_1, T_2, T_3)$. Podemos notar facilmente que o número de maneiras possíveis de formar

blocos contíguos com os elementos do vetor $\mathbf{T}$ é igual a 2^{n-1}, em que neste caso $n = 3$.

Tabela 3.1: Diferentes blocos contíguos formados para T_1, T_2, T_3.

Blocos	Nº de blocos	ρ
$[T_1][T_2][T_3]$	3	$\{0, 1, 2, 3\}$
$[T_1, T_2][T_3]$	2	$\{0, 2, 3\}$
$[T_1][T_2, T_3]$	2	$\{0, 1, 3\}$
$[T_1, T_2, T_3]$	1	$\{0, 3\}$

Segundo Barry & Hartigan (1992), dizemos que $(T_1, T_2 \ldots, T_n, \rho) \sim MPP$ se:

1. A distribuição *a priori* de ρ assume a seguinte forma produto

$$p(\rho) = K^{-1} \prod_{j=1}^{b} c_{[i_{j-1}, i_{j-1}]},$$

em que K é uma constante definida como a soma sobre todas as partições seja igual a 1, e $c_{[i_{j-1}, i_{j-1}]}$ é denominada de coesão *a priori*, que mede o grau de similaridade das observações que estão sendo agrupadas no bloco $[i_{j-1}; i_j]$;

2. Condicional em ρ, a função densidade conjunta é dada por

$$f(\mathbf{T} \mid \rho) = \prod_{j=1}^{b} f\left(\mathbf{T}_{[i_{j-1}, i_{j-1}]}\right).$$

Sob um enfoque paramétrico, denotamos a função densidade por $f(\mathbf{T} \mid \boldsymbol{\theta})$, com $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_n\}$. Assumimos que, $\theta_1, \ldots, \theta_n$, $T_1, \ldots, T_n$ são independentes. Assim, temos que a função densidade conjunta é dada por

$$f(\mathbf{T} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} f(T_i \mid \theta_i),$$

em que $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$. A construção da distribuição *a priori* para $\theta_1, \theta_2, \ldots, \theta_n$ é feita da seguinte forma:

1. Dada a partição aleatória ρ, assumimos que a sequência de parâmetros $\theta_1, \theta_2, \ldots, \theta_n$ pode ser dividida em b blocos contíguos, em que θ_i são parâmetros comuns dentro de cada bloco, ou seja,

$$\theta_i = \theta_{[i_{j-1}, i_j]}, \quad \forall\ i_{j-1} < i \leq i_j,$$

2. Assumimos que os parâmetros comuns $\theta_{i_0, i_1}, \theta_{i_1, i_2}, \ldots, \theta_{i_{b-1}, i_b}$ são independentes e seguem uma distribuição *a priori* $p_{[i_{j-1}, i_{j-1}]}(\theta_{[i_{j-1}, i_{j-1}]})$ para $j = 1, \ldots, b$.

A distribuição conjunta das observações e dos parâmetros de interesse é dada por

$$f(\mathbf{T};\theta \mid \rho) = \prod_{j=1}^{b} f(\mathbf{T}_{[i_{j-1},i_{j-1}]};\theta_{[i_{j-1},i_{j-1}]}),$$

em que

$$f(\mathbf{T}_{[i_{j-1},i_{j-1}]};\theta_{[i_{j-1},i_{j-1}]}) = \prod_{r=i_{j-1}+1}^{i_j} f(T_r \mid \theta_{[i_{j-1},i_{j-1}]}) p_{[i_{j-1},i_{j-1}]}(\theta_{[i_{j-1},i_{j-1}]}),$$

em que $f(\mathbf{T}_{[i_{j-1},i_{j-1}]};\theta_{[i_{j-1},i_{j-1}]})$ representa a função de densidade conjunta para da bloco $[i_{j-1}, i_{j-1}]$.

A distribuição preditiva associada ao bloco $[i_{j-1}; i_j]$ é expressa por

$$f(\mathbf{T}_{[i_{j-1},i_{j-1}]}) = \prod_{r=i_{j-1}+1}^{i_j} \int_{\boldsymbol{\theta}} f(T_k \mid \theta_{[i_{j-1},i_{j-1}]}) p_{[i_{j-1},i_{j-1}]}(\theta_{[i_{j-1},i_{j-1}]}) \quad d\theta_{[i_{j-1},i_{j-1}]}.$$

Consequentemente, a distribuição *a posteriori* para a partição aleatória é dada por

$$p(\rho \mid \mathbf{T}) \propto \prod_{j=1}^{b} f(\mathbf{T}_{[i_{j-1},i_{j-1}]})\ c_{[i_{j-1},i_{j-1}]}$$

Para amostrarmos da distribuição *a posteriori* de ρ, utilizamos o algoritmo amostrador de Gibbs proposto por Barry & Hartigan (1993).

Para realizarmos inferência para $\theta_1, \theta_2, \ldots, \theta_n$, definimos inicialmente a distribuição *a posteriori* para cada bloco dos parâmetros, tendo como base as observações dispostas em cada bloco, pela seguinte expressão

$$p_{[i_{j-1},i_{j-1}]}(\theta_{[i_{j-1},i_j]} \mid \mathbf{T}_{[i_{j-1},i_j]}) \propto f(\mathbf{T}_{[i_{j-1},i_j]} \mid \theta_{[i_{j-1},i_j]}) p_{[i_{j-1},i_{j-1}]}(\theta_{[i_{j-1},i_j]}).$$

Então, condicional em $\mathbf{T}$, a distribuição *a posteriori* de θ_r, para $r = 1, \ldots, n$, é obtida pela mistura das densidades *a posteriori* associados aos diferentes blocos é dada por

$$p(\theta_r \mid D) = \sum_{i_{j-1}<r<i_j} p(\theta_k \mid \mathbf{T}_{[i_{j-1},i_{j-1}]}) R([i_{j-1}; i_j])$$

em que quantidade $R([i_{j-1}; i_j])$ representa a probabilidade *a posteriori* do bloco $[i_{j-1}; i_j]$ aparecer na partição $\rho = \{i_0, i_1, ..., i_b\}$ denominada de relevância *a posteriori* definida pela seguinte expressão

$$R([i_{j-1}; i_j]) = \frac{a_{i_0 i_{j-1}} c^*_{i_{j-1} i_j} a_{i_j i_b}}{a_{i_0 i_b}}$$

com $a_{i_{j-1} i_j} = \sum_{i=1}^{n} \prod_{j=1}^{b} c^*_{i_{j-1} i_j}$ e $c^*_{i_{j-1} i_j} = f(\mathbf{T}_{[i_{j-1},i_j]}) c_{[i_{j-1},i_j]}$. O cálculo direto da relevância *a posteriori* requer um grande esforço computacional. Dessa forma, utilizaremos o algoritmo

amostrador de Gibbs proposto por Barry & Hartigan (1993) que dispensam o cálculo das relevâncias.

3.1.1 Algoritmo de Barry & Hartigan (1993)

Apresentamos nesta seção o algoritmo amostrador de Gibbs, proposto por Barry & Hartigan (1993), para a atualização da partição aleatória ρ. O algoritmo desenvolvido por Barry & Hartigan (1993) foi proposto para determinar blocos de observações, ou seja, de agrupar uma sequencia de de observações dispostas no tempo. Outra característica do algoritmo é que as relevâncias a *a posteriori* podem ser calculas numericamente.

Seja $\mathbf{E} = (E_1, E_2, \ldots, E_{n-1})$ uma sequência de variáveis aleatórias tais que

$$E_i = \begin{cases} 1, & \theta_i = \theta_{i+1} \\ 0, & \theta_i \neq \theta_{i+1}, \end{cases} \quad j = 1, \ldots, n-1.$$

Quando $E_i = 1$ os blocos $[i_{j-1}; i_j]$ e $[i_j; i_{j+1}]$ associados a $\theta_{[i_{j-1};i_j]}$ e $\theta_{[i_j;i_{j+1}]}$, respectivamente, são denotados por $[i_{j-1}; i_{j+1}]$, com taxa $\theta_{[i_{j-1};i_{j+1}]}$ e quando $E_i = 0$ os blocos $[i_{j-1}; i_j]$ e $[i_j; i_{j+1}]$ são separados. Note que a partição aleatória $\rho = \{i_0, i_1, ..., i_b\}$ pode ser representada por $\mathbf{E} = \{E_1, E_2, \ldots, E_{n-1}\}$, em que cada componente de $\mathbf{E}$ nos indica quais são os pontos de mudança em ρ.

O algoritmo para a atualização de ρ segue os seguintes passos:

1. Inicie com $s = 0$ o vetor $\mathbf{E}^{(s)} = (E_1^{(s)}, E_2^{(s)}, \ldots, E_{n-1}^{(s)})$, com $E_i^{(0)} = 0$ para $i = 1, \ldots, n-1$.

2. Para o atualizar a $i-$ésima componente de $\mathbf{E}^{(s)}$, condicional em todas as demais componentes, ou seja,

$$E_i^{(s)} \mid E_1^{(s)}, E_2^{(s)}, \ldots, E_{i-1}^{(s)}, E_{i+1}^{(s-1)}, \ldots, E_{n-1}^{(s-1)}, i = 1, \ldots, n-1,$$

faça a seguinte razão

$$R_i^{(s)} = \frac{P(E_i^{(s)} = 1 \mid \mathbf{E}_{-i}, D)}{P(E_i^{(s)} = 0 \mid \mathbf{E}_{-i}, D)} = \frac{\prod_{j=1}^{b-1} f(\mathbf{T}_{[i_{j-1},i_{j-1}]})\ c_{[i_{j-1},i_{j-1}]}}{\prod_{j=1}^{b} f(\mathbf{T}_{[i_{j-1},i_{j-1}]})\ c_{[i_{j-1},i_{j-1}]}}, i = 1, \ldots, n-1,$$

em que $\mathbf{E}_{-i} = \left(E_1^{(s)}, E_2^{(s)}, \ldots, E_{i-1}^{(s)}, E_{i+1}^{(s-1)}, \ldots, E_{n-1}^{(s-1)}\right)$. Assim,

$$E_i^{(s)} = \begin{cases} 1, & R_i^{(s)} \geq u/1-u \\ 0, & cc, \end{cases}, i = 1, \ldots, n-1$$

em que $u \sim Unif[0,1]$.

3. Incremente $s+1$, e volte ao passo 2, até a convergência da cadeia.

A seguir descreveremos como utilizar a estrutura de agrupamento do MPP para modelar a grade dos tempos do MEP.

3.2 MEP com grade aleatória

Denote por $\mathscr{F} = \{0, t_1, ..., t_m\}$ o conjunto formado pela origem e m tempos de falha distintos e ordenados de uma amostra de tamanho n. Considere $\tau' = \{0, t'_1, ..., t'_{m'-1}, t'_{m'}\}$, como sendo uma grade de tempos inicial, satisfazendo $\tau' \subseteq \mathscr{F}$, para, $1 \leq m' \leq m$. Dessa forma, τ' induz o seguinte conjunto de intervalos

$$I_j = \begin{cases} (0, t'_1], \text{ se } j = 1, \\ (t'_{j-1}, t'_j], \text{ para } j = 2, 3, \ldots, m'. \end{cases}$$

A abordagem proposta por Demarqui et al. (2008) consiste em especificar uma grade inicial τ' admitida *a priori* e usar os dados para agrupar os intervalos induzidos por τ'. Denote por $\mathcal{I} = \{1, ..., m'\}$ o conjunto de índices associados a $I_1, ..., I_{m'}$ e considere $\rho = \{i_0, i_1, ..., i_b\}$, $0 = i_0 < i_1 < ... < i_b = m'$, uma partição aleatória de $\mathcal{I}$ que divide os m' intervalos iniciais em b novos intervalos aleatórios contíguos e disjuntos. Defina $\tau = \tau(\rho) = \{s_0, s_1, ..., s_b\}$ uma grade aleatória tal que

$$s_j = \begin{cases} 0, \text{ se } j = 0, \\ t'_{i_j}, \text{ se } j = 1, \ldots, b. \end{cases}$$

Observe que os intervalos agrupados induzidos por $\rho = \{i_0, i_1, ..., i_b\}$ são da forma

$$I_\rho^{(j)} = \cup_{r=i_{j-1}+1}^{i_j} I_r = (s_{j-1}, s_j], \quad j = 1, ..., b.$$

Então, dado $\rho = \{i_0, i_1, ..., i_b\}$, assumimos que a função risco é dada por

$$h(t) = \lambda_r \equiv \lambda_\rho^{(j)},$$

em que $\lambda_\rho^{(j)}$ denota a taxa comum associada ao intervalo agrupado $I_\rho^{(j)}$, para $i_{j-1} < r \leq i_j$, com $r = 1, ..., m'$ e $j = 1, ..., b$. A equivalência acima estabelece uma relação entre a taxa da grade inicial e as taxas de falha referentes aos intervalos agrupados. A Figura 3.1 fornece uma ilustração da estrutura de agrupamento proposta por Demarqui et al. (2008).

$$\tau' = \{0, t_1', t_2', t_3', t_4', t_5'\} \quad \leftrightarrow \quad \mathcal{I} = \{1, 2, 3, 4, 5\}$$

$$\downarrow$$

$$\underbrace{I_1 = (0, t_1'];}_{\lambda_1} \quad \underbrace{I_2 = (t_1', t_2'];}_{\lambda_2} \quad \underbrace{I_3 = (t_2', t_3'];}_{\lambda_3} \quad \underbrace{I_4 = (t_3', t_4'];}_{\lambda_4} \quad \underbrace{I_5 = (t_4', t_5']}_{\lambda_5}$$

$$\downarrow$$

$$\mathcal{I} = \{1, 2, 3, 4, 5\} \quad \rightarrow \quad \rho = \{i_0 = 0, i_1 = 3, i_2 = 5\}$$

$$\downarrow$$

$$\underbrace{I_\rho^{(1)} = (0, t_3'];}_{\lambda_\rho^{(1)}} \quad \underbrace{I_\rho^{(2)} = (t_3', t_5']}_{\lambda_\rho^{(2)}}$$

Figura 3.1: Ilustração da estrutura de agrupamento do MPP , com $m' = 5$ e $b = 2$.

Suponha uma amostra de tamanho n em que os indivíduos estão sujeitos a censura à direita, em que $y_i = \min(t_i, c_i)$ é o valor observado. Para construirmos a função de verossimilhança sobre os b intervalos induzidos por $\tau(\rho) = \{s_0, s_1, ..., s_b\}$, é conveniente definirmos

$$y_{ij} = \begin{cases} s_{j-1}, & \text{se } y_i < s_{j-1}, \\ y_i, & \text{se } y_i \in I_\rho^{(j)}, \\ s_j, & \text{se } y_i > s_j, \end{cases}$$

e

$$\delta_{ij} = \begin{cases} 1, & \text{se } y_i \in I_\rho^{(j)} \text{ e } \delta_i{=}1, \\ 0, & \text{caso contrário,} \end{cases}$$

em que $I_\rho^{(j)} = (s_{j-1}, s_j]$, com $j = 1, ..., b$ e $i = 1, ..., n$. Então, utilizando (2.8), a função de verossimilhança pode ser expressa da seguinte forma

$$\begin{aligned} L(\boldsymbol{\lambda}_\rho, \rho | D) &= \prod_{i=1}^{n} \left[\prod_{j=1}^{b} \left(\lambda_\rho^{(j)}\right)^{\delta_{ij}} \exp\left\{-\lambda_\rho^{(j)}(y_{ij} - s_{j-1})\right\} \right] \\ &= \prod_{j=1}^{b} \left(\lambda_\rho^{(j)}\right)^{\nu_j} \exp\left(-\lambda_\rho^{(j)} \xi_j\right), \end{aligned} \tag{3.5}$$

em que $D = \{(y_i, \delta_i) : i = 1, \ldots, n\}$, $\nu_j = \sum_{i=1}^{n} \delta_{ij}$ e $\xi_j = \sum_{i=1}^{n} (y_{ij} - s_{j-1})$, $j = 1, ..., b$.
Observe que a verossimilhança dada em (3.5) fatora em um produto de núcleos de distribuições gama, permitindo assim que a estrutura de agrupamento do modelo partição produto proposto por Barry & Hartigan (1992) seja utilizada para modelar a grade τ (Demarqui

et al., 2008).

3.3 Distribuições *a priori*

A distribuição *a priori* conjunta para $(\boldsymbol{\lambda}_\rho, \rho)$ é dada por

$$p(\boldsymbol{\lambda}_\rho, \rho) = p(\boldsymbol{\lambda}_\rho \mid \rho)p(\rho). \tag{3.6}$$

É importante observar que há uma estrutura hierárquica entre $\boldsymbol{\lambda}_\rho$ e ρ. Dessa forma, apresentamos primeiramente a distribuição *a priori* para a partição aleatória ρ, e, em seguida, a distribuição *a priori* para $\boldsymbol{\lambda}_\rho$, condicional em ρ.

A distribuição *a priori* para a partição aleatória ρ é dada pela seguinte forma produto

$$p(\rho = \{i_0, i_1, \ldots, i_b\}) = K^{-1} \prod_{j=1}^{b} c_{I_\rho^{(j)}}, \tag{3.7}$$

em que $0 = i_0 < i_1 < \cdots < i_b = m^{'}$, $b \in \mathcal{I}$ e $c_{I_\rho^{(j)}}$ é um valor positivo denominado de coesão *a priori*, que representa o grau de similaridade entre os intervalos associados a grade inicial que serão agrupados. A coesão *a priori* é uma quantidade muito importante, pois é através dela que induzimos a forma de agrupamento dos intervalos, ou seja, se temos algum conhecimento prévio acerca do agrupamento dos intervalos, essa informação pode ser representada por $c_{I_\rho^{(j)}}$. A constante é definida como $K = \sum_{\mathscr{C}} \prod_{j=1}^{b} c_{I_\rho^{(j)}}$.

Note que o MEP com grade fixa corresponde ao caso particular do MEP com grade aleatória quando $p(\rho = \{i_0, i_1, \ldots, i_b\}) = 1$ para uma particular partição.

A distribuição *a priori* de $(\boldsymbol{\lambda}_\rho \mid \rho)$ é expressa na forma de produto, ou seja,

$$p(\boldsymbol{\lambda}_\rho \mid \rho) \;\; \propto \;\; \prod_{j=1}^{b} p(\lambda_\rho^{(j)}).$$

Para a distribuição *a priori* de $\lambda_\rho^{(j)}$, assumimos uma distribuição gama com parâmetros de forma e escala α_j e γ_j, com $j = 1, \ldots, b$. A distribuição gama é rica em formas e corresponde à distribuição *a priori* conjugada para as taxas de falha $\lambda_\rho^{(j)}$ que compõem a função de verossimilhança em (3.5), permitindo dessa forma utilizar a estrutura de agrupamento do MPP para modelar a grade do MEP.

Assim, substituindo os elementos da distribuição *a priori* conjunta de $(\boldsymbol{\lambda}_\rho, \rho)$ em (3.6), temos

$$p(\boldsymbol{\lambda}_\rho, \rho) \propto \prod_{j=1}^{b} \left\{ \left(\lambda_\rho^{(j)}\right)^{\alpha_j - 1} \exp\left(-\gamma_j \lambda_\rho^{(j)}\right) c_{I_\rho^{(j)}} \right\}. \tag{3.8}$$

Através da função de verossimilhança em (3.5), e da distribuição *a priori* em (3.8), a

distribuição dos dados, condicional em ρ, também segue uma forma produto, isto é,

$$\begin{aligned} f(D \mid \rho) &= \int \prod_{j=1}^{b} L(\boldsymbol{\lambda}_\rho, \rho|D)\ p(\lambda_\rho^{(j)})\ d\lambda_\rho^{(j)} \\ &= \prod_{j=1}^{b} \int \left(\lambda_\rho^{(j)}\right)^{\nu_j} \exp\left(-\lambda_\rho^{(j)}\xi_j\right)\ \frac{\gamma_j^{\alpha_j}}{\Gamma(\alpha_j)} \left(\lambda_\rho^{(j)}\right)^{\alpha_j-1} \exp\left(-\gamma_j\lambda_\rho^{(j)}\right)\ d\lambda_\rho^{(j)} \\ &= \prod_{j=1}^{b} \int \frac{\gamma_j^{\alpha_j}}{\Gamma(\alpha_j)} \left(\lambda_\rho^{(j)}\right)^{\nu_j+\alpha_j-1}\ \exp\left\{-\lambda_\rho^{(j)}(\xi_j+\gamma_j)\right\}\ d\lambda_\rho^{(j)} \\ &= \prod_{j=1}^{b} \frac{\gamma_j^{\alpha_j}}{\Gamma(\alpha_j)}\ \frac{\Gamma(\nu_j+\alpha_j)}{(\xi_j+\gamma_j)^{\nu_j+\alpha_j}}. \end{aligned} \tag{3.9}$$

3.3.1 Distribuição *a posteriori*

A distribuição *a posteriori* conjunta de $(\boldsymbol{\lambda}_\rho, \rho)$ é escrita da seguinte forma:

$$p(\boldsymbol{\lambda}_\rho, \rho \mid D) \propto \prod_{j=1}^{b} \frac{\gamma_j^{\alpha_j}}{\Gamma(\alpha_j)} \left(\lambda_\rho^{(j)}\right)^{\nu_j+\alpha_j-1} \exp\left\{-\lambda_\rho^{(j)}(\xi_j+\gamma_j)\right\} c_{I_\rho^{(j)}}. \tag{3.10}$$

Ao observar a distribuição *a posteriori* em (3.10), notamos que a mesma não tem uma forma analítica conhecida. Portanto, métodos MCMC devem ser utilizados para se obter uma amostra de $(\boldsymbol{\lambda}_\rho,\ \rho)$.

Utilizando a distribuição dos dados em (3.9) e a distribuição *a priori* de ρ em (3.7), a distribuição *a posteriori* para ρ é dada por

$$\begin{aligned} p(\rho \mid D) &\propto f(D \mid \rho)p(\rho) \\ &\propto \prod_{j=1}^{b} \frac{\gamma_j^{\alpha_j}}{\Gamma(\alpha_j)} \frac{\Gamma(\nu_j+\alpha_j)}{(\xi_j+\gamma_j)^{\nu_j+\alpha_j}} c_{I_\rho^{(j)}}. \end{aligned} \tag{3.11}$$

Para amostrar da distribuição *a posteriori* da partição aleatória ρ, será utilizado o amostrador de Gibbs proposto em Barry & Hartigan (1993).

Condicionado à partição aleatória ρ, a distribuição condicional completa para as taxas de cada intervalo agrupado induzido por ρ é escrita da seguinte forma:

$$p(\boldsymbol{\lambda}_\rho \mid \rho, D) \propto \prod_{j=1}^{b} \left(\lambda_\rho^{(j)}\right)^{\nu_j+\alpha_j-1} \exp\left\{-\lambda_\rho^{(j)}(\xi_j+\gamma_j)\right\}. \tag{3.12}$$

Observe que $\left(\lambda_\rho^{(j)} \mid \rho, D\right) \sim \text{Gama}(\nu_j+\alpha_j, \xi_j+\gamma_j)$, para $j = 1, \ldots, b$. Portanto, amostras de $\lambda_\rho^{(j)}$ podem ser geradas de forma direta. Dessa forma, seguindo a estrutura do MPP, a distribuição *a posteriori* para λ_k, com $k = 1, 2, \ldots, m'$, ou seja, a distribuição da taxa de

falha associada à grade inicial, é determinada pela seguinte mistura de distribuições,

$$p(\lambda_k \mid D) = \sum_{i_{j-1}<k<i_j} p(\lambda_\rho^{(j)} \mid D) R(I_\rho^{(j)}), \tag{3.13}$$

em que $R(I_\rho^{(j)})$ denota a relevância *a posteriori*, e aqui representa a probabilidade de cada intervalo agrupado $I_\rho^{(j)}$ pertencer à grade aleatória induzida pela partição aleatória ρ, e $p(\lambda_\rho^{(j)} \mid D)$ representa a distribuição *a posteriori* com taxa $\lambda_\rho^{(j)}$ comum, com $j = 1, 2, \ldots, b$.

A função de sobrevivência *a posteriori*, baseada no MEP com grade aleatória é dada por

$$S(t \mid D) = \sum_{\rho} S(t \mid D, \rho) p(\rho \mid D), \tag{3.14}$$

em que

$$\begin{aligned} S(t \mid D, \rho) &= \int S(t \mid \boldsymbol{\lambda}_\rho, \rho)\ p(\boldsymbol{\lambda}_\rho)\ d\boldsymbol{\lambda}_\rho \\ &= \left(1 + \frac{t - s_{j-1}}{\gamma_j}\right)^{\alpha_j} \prod_{r=1}^{j-1} \left(1 + \frac{s_r - s_{r-1}}{\gamma_r}\right)^{\alpha_r}, \end{aligned} \tag{3.15}$$

com $t \in I_\rho^{(j)}$, com $j = 1, \ldots, b$.

Capítulo 4

Modelos semiparamétricos para dados de sobrevivência com censura intervalar

Neste capítulo apresentamos as contribuições desta tese para a modelagem de dados de sobrevivência com censura intervalar. Para o caso sem fração de cura, descrevemos os modelos propostos por Sinha et al. (1999) e Wang et al. (2013) com intuito de comparação. A seguir, apresentamos a adaptação do modelo de Demarqui et al. (2012) para dados de sobrevivência com censura intervalar. Com base na estrutura dinâmica apresentada em Demarqui (2010), apresentamos um novo modelo dinâmico semiparamétrico para dados de sobrevivência com censura intervalar e com a presença de uma fração de curados.

4.1 Modelos de sobrevivência sem fração de cura

Nesta seção, apresentamos os modelos a serem aplicados a dados de sobrevivência sujeitos à censura intervalar. Inicialmente iremos mostrar dois modelos já disponíveis na literatura para dados de sobrevivência com censura intervalar, os quais iremos comparar em seguida com os modelos propostos nesta tese.

4.1.1 Modelos de Sinha et al. (1999)

Sinha et al. (1999) propuseram uma abordagem Bayesiana semiparametrica para a modelagem de dados de sobrevivência sujeitos à censura intervalar em que a função risco de base é modelada pelo MEP com grade fixa. No modelo proposto por estes autores, é assumido ainda um vetor de coeficientes da regressão diferente para cada intervalo da grade de tempos que define o MEP, permitindo desta forma que os efeitos das covariáveis variem no tempo.

A função risco do modelo proposto por Sinha et al. (1999) é dada por

$$h(t|\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\beta}_j) = \lambda_j \exp(\boldsymbol{x}^\top \boldsymbol{\beta}_j), \quad \text{se} \quad t \in I_j, \ j = 1, 2, \ldots, b, \tag{4.1}$$

em que $I_j = (s_{j-1}, s_j]$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_j^\top : j = 1, 2, \ldots, b)$, com $\boldsymbol{\beta}_j = (\beta_{j1}, \beta_{j2}, \ldots, \beta_{jp})^\top$, dessa forma,

$\boldsymbol{\beta}_j$ é o mesmo em cada intervalo de tempo e $0 = s_0 < s_1 < \cdots < s_b = \infty$.

Observe que a equação (4.1) é uma extensão ao modelo de riscos proporcionais, uma vez que utiliza uma estrutura que permite o vetor dos coeficientes da regressão variar no tempo e, assim, a propriedade de riscos proporcionais não é mais satisfeita, uma vez que a razão dos riscos entre o $i-$ésimo e $j-$ésimo indivíduo depende do tempo.

Seja $D_c = \{(y_i, \delta_i, \xi_i) : i = 1, \ldots, n\}$ o conjunto dos dados completos. Logo, a função de verossimilhança condicional em D_c, é dada por

$$L(\boldsymbol{\lambda}, \boldsymbol{\beta}|D_c) = \prod_{j=1}^{b} \exp\left(\sum_{i=1}^{n} \delta_{ij} \boldsymbol{x}_i^\top \boldsymbol{\beta}_j\right) (\lambda_j)^{\nu_j} \exp\left(-\lambda_j \xi_j\right), \tag{4.2}$$

em que $\nu_j = \sum_{i=1}^{n} \delta_{ij}$ é o número total de falhas no j-ésimo intervalo e

$$\xi_j = \sum_{i=1}^{n} \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j\right) (y_{ij} - s_{j-1}),$$

com y_{ij} definido como em (3.2).

A distribuição *a priori* conjunta para $\boldsymbol{\lambda}$ e $\boldsymbol{\beta}$ é dada por

$$p(\boldsymbol{\lambda}, \boldsymbol{\beta}) = p(\boldsymbol{\lambda})\ p(\boldsymbol{\beta}) = \prod_{j=1}^{b} \left[p(\lambda_j)\ p(\boldsymbol{\beta}_j|\boldsymbol{\beta}_{j-1})\right].$$

Observe que os componentes do vetor $\boldsymbol{\beta}$ são correlacionados. Para as taxas do MEP, assume-se que os $\lambda_j's$ são independentes *a priori* e $\lambda_j \sim \text{Gama}(\alpha_j, \gamma_j)$, com $j = 1, 2, \ldots, b$. A estrutura dinâmica que acomoda os coeficientes dependendo do tempo é da forma

$$\boldsymbol{\beta}_j = \boldsymbol{\beta}_{j-1} + \mathbf{Q}_j, \quad \mathbf{Q}_j \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}_j),$$

em que $\boldsymbol{\Sigma}_j = \text{diag}(\omega_{1j}^2, \omega_{2j}^2, \ldots, \omega_{pj}^2)$. Então, $\boldsymbol{\beta}_j|\boldsymbol{\beta}_{j-1} \sim N_p(\boldsymbol{\beta}_{j-1}, \boldsymbol{\Sigma}_j)$ e por simplicidade tomamos $p(\boldsymbol{\beta}_j|\boldsymbol{\beta}_{j-1}) = \prod_{k=1}^{p} f(\beta_{k(j)}|\beta_{k(j-1)}, w_{kj}^2)$, em que $f(\beta_{k(j)}|\beta_{k(j-1)}, w_{kj}^2)$ é a função densidade de uma distribuição normal com média $\beta_{k(j-1)}$ e variância w_{kj}^2, para $j = 1, ..., b$.

A distribuição *a posteriori*, conjunta condicional nos dados completos, é dada por

$$\begin{aligned} p(\boldsymbol{\lambda}, \boldsymbol{\beta}|D_c) &\propto L(\boldsymbol{\lambda}, \boldsymbol{\beta}|D_c)\ p(\boldsymbol{\lambda}, \boldsymbol{\beta}) \\ &\propto \prod_{j=1}^{b} \exp\left(\sum_{i=1}^{n} \delta_{ij} \boldsymbol{x}_i^\top \boldsymbol{\beta}_j\right) (\lambda_j)^{\nu_j} \exp\left(-\lambda_j \xi_j\right)\ p(\lambda_j)\ p(\boldsymbol{\beta}_j|\boldsymbol{\beta}_{j-1}). \end{aligned} \tag{4.3}$$

Observe que a distribuição *a posteriori* conjunta acima não apresenta forma analítica conhecida. Neste caso, precisamos de algum método de aproximação para a obtenção das estimativas de $\boldsymbol{\lambda}$ e $\boldsymbol{\beta}$. No amostrador de Gibbs, as distribuições condicionais completas baseadas

nos dados completos são

$$[\lambda_j|\boldsymbol{\beta}, D_c] \sim \text{Gama}(\alpha_j + \nu_j, \gamma_j + \xi_j),$$

e

$$[\boldsymbol{\beta}_j|\boldsymbol{\lambda}, \boldsymbol{\beta}_j^{(-r)}, D_c] \propto \prod_{j=1}^{b} e^{\sum_{i=1}^{n} \delta_{ij} \boldsymbol{x}_i^\top \boldsymbol{\beta}_j} \exp\left(-\lambda_j \xi_j\right) p(\boldsymbol{\beta}_j|\boldsymbol{\beta}_{j-1}),$$

em que $\boldsymbol{\beta}_j^{(-r)}$, representa o vetor $\boldsymbol{\beta}_j$ sem a $r-$ésima componente. Fixado o intervalo j, o método ARS pode ser utilizado para gerar de cada componente de $\boldsymbol{\beta}_j$, pois $p(\boldsymbol{\beta}_j|\boldsymbol{\lambda}, \boldsymbol{\beta}_j^{(-r)}, D_c)$ é log-côncava em cada componente do vetor $\boldsymbol{\beta}_j$.

4.1.2 O modelo de Wang et al. (2013)

Uma extensão ao modelo proposto por Sinha et al. (1999) é o modelo proposto por Wang et al. (2013). Segundo Wang et al. (2013), fazer o uso de distribuições gamas independentes para as taxas do MEP pode acarretar em uma sobreparametrização do modelo e, desta forma, as estimativas *a posteriori* dos parâmetros apresentarão grandes variações. Outro problema consiste em termos que especificar a variância ω_j^2 para cada intervalo, ao assumir uma distribuição normal para cada componente do vetor de coeficientes de regressão. Este problema se torna mais acentuado quando o número de intervalos é grande, podendo ocasionar instabilidade numérica. Além disso, em Wang et al. (2013) a grade τ é tratada como um parâmetro adicional a ser estimado.

Uma solução a esses problemas, proposta por Wang et al. (2013), consiste em modelar simultaneamente $\beta_{0j} = \log(\lambda_j)$ (como um intercepto aleatório) e $\boldsymbol{\beta}_{kj}$, para $k = 1, \ldots, p$ e $j = 1, \ldots, b$, através de uma função $\theta(t)$ que combina ambas quantidades, ou seja, utilizamos $\theta(t)$ como um vetor de contém $\beta_{0j} = \log(\lambda_j)$ e os outros componentes $\boldsymbol{\beta}_{kj}$. Nesta abordagem é utilizado um processo Markoviano *a priori* para $\theta(t)$.

As especificações *a priori* para a construção do modelo são as seguintes:

1. Assuma que o número de saltos b em $\theta(t)$ siga uma distribuição uniforme discreta com amplitude de 1 a $m^{'}$, em que $m^{'}$ é número máximo de intervalos. Dado b, os tempos de saltos são determinados por $0 < s_1 < s_2 < \cdots < s_b$, em que somente o último ponto de salto não pode ser selecionado de maneira aleatória. Note que, assim como no MEP via MPP, no modelo proposto por Wang et al. (2013) também é necessário especificar um grade inicial, que pode ser construída a partir dos limites observados e distintos l_i e r_i ou uma grade de tempos equiespaçados.

2. A distribuição *a priori* baseada no processo Markoviano para $\theta(t)$ é dada por

$$\begin{aligned} \theta(s_1) \mid \omega^2 &\sim N(0, a_0\omega^2), \\ \theta(s_j) \mid \theta(s_{j-1}), \omega^2 &\sim N(\theta(s_{j-1}), \omega^2), j = 2, \ldots, b, \\ \omega^2 &\sim GI(\alpha_0, \eta_0), \end{aligned} \tag{4.4}$$

em que a_0 é um hiperparâmetro ($a_0 > 0$) que controla a variabilidade no primeiro intervalo, $GI(\alpha_0, \eta_0)$ denota uma distribuição gama inversa com parâmetros de forma $\alpha_0 > 0$ e escala $\eta_0 > 0$, com média $\eta_0/(\alpha_0 - 1)$ e variância $(\eta_0/(\alpha_0 - 1)^2(\alpha_0 - 2))$, para $\alpha > 2$. Dessa forma, a distribuição *a priori* conjunta é escrita como segue:

$$\begin{aligned} p(\theta(t), \omega^2) \propto\ & \frac{{\eta_0}^{\alpha_0}}{\Gamma(\alpha_0)}(\omega^2)^{-\alpha_0-1}\exp\left(\frac{\eta_0}{\omega^2}\right)(\omega^2)^{-\frac{b}{2}}\exp\left\{-\frac{\theta(s_1)^2}{2a_0\omega^2}\right\} \\ & \times\prod_{j=2}^{b}\exp\left\{-\frac{(\theta(s_j)-\theta(s_{j-1}))^2}{2\omega^2}\right\}. \end{aligned}$$

A distribuição *a posteriori* conjunta, condicional nos dados completos, é dada por

$$\begin{aligned} p(\boldsymbol{\beta}, \omega^2, b \mid D_c) &\propto L(\boldsymbol{\lambda}, \boldsymbol{\beta}|D_c)\ p(\theta(t), \omega^2) \\ &\propto \prod_{j=1}^{b}\exp\left(\sum_{i=1}^{n}\delta_{ij}\boldsymbol{x}_i^\top\boldsymbol{\beta}_j\right)(\lambda_j)^{\nu_j}\exp\left(-\lambda_j\xi_j\right) \\ &\quad\times\frac{{\eta_0}^{\alpha_0}}{\Gamma(\alpha_0)}(\omega^2)^{-\alpha_0-1}\exp\left(\frac{\eta_0}{\omega^2}\right)(\omega^2)^{-\frac{b}{2}}\exp\left\{-\frac{\theta(s_1)^2}{2a_0\omega^2}\right\} \\ &\quad\times\prod_{j=2}^{b}\exp\left\{-\frac{(\theta(s_j)-\theta(s_{j-1}))^2}{2\omega^2}\right\}. \end{aligned} \tag{4.5}$$

Observe que a distribuição *a posteriori* acima permite que a dimensão do espaço paramétrico varie, uma vez que o número de saltos b é aleatório, tornando necessário o uso do algoritmo de Monte Carlo via cadeias de Markov com saltos reversíveis (Green, 1995), para amostrar da distribuição *a posteriori* em (4.5). A descrição completa do algoritmo RJMCMC para este modelo pode ser encontrada em Wang et al. (2013) ou em Chen et al. (2013) (no Capítulo 7).

4.1.3 Modelo dinâmico com efeito variando no tempo

O modelo de Gamerman (1991) estende o modelo de riscos proporcionais proposto por Cox (1972), admitindo que os coeficientes das covariáveis variem em intervalos de tempo, através de um sistema de equações que fornecem uma descrição da evolução estocástica dos parâmetros ao longo do tempo. Tal abordagem utiliza uma análise sequencial, baseada na fatoração da função de verossimilhança nos intervalos de tempo, assumindo a distribuição exponencial por partes para os tempos de falha. Em Demarqui et al. (2012) é apresentada uma extensão do modelo proposto por Gamerman (1991) em que a grade τ é considerada uma quantidade aleatória.

Apresentamos a seguir uma extensão do modelo proposto por Demarqui et al. (2012) para acomodar dados de sobrevivência sujeitos a censura intervalar, condicionando nos pseudo-tempos de falha gerados para cada intervalo observado. Observe que, condicional nos dados ampliados, os modelos de Demarqui et al. (2012) e Gamerman (1991) podem ser utilizados.

O desenvolvimento a seguir é realizado condicionando-se nos dados ampliados e na partição aleatória ρ, resultando em uma descrição da abordagem proposta por Gamerman (1991). Então, seja T é uma variável aleatória representando o tempo de falha, tal que $T \sim MEP(\boldsymbol{\lambda}_\rho, \rho)$, em que a função de taxa de falha é

$$h(t \mid \boldsymbol{x}, \boldsymbol{\lambda}_\rho, \rho) = \lambda_\rho^{(j)} = \exp(\boldsymbol{x}^\top \boldsymbol{\beta}_\rho^{(j)}), \tag{4.6}$$

em que $t \in I_\rho^{(j)} = (s_{j-1}, s_j]$, $j = 1, \ldots, b$ e que $\boldsymbol{x} = (1, x_1, \ldots, x_p)^\top$, com x_k, $k = 1, \ldots, p$, sendo o valor da $k-$ésima covariável para o $i-$ésimo indivíduo e $\boldsymbol{\beta}_\rho^{(j)'} = (\boldsymbol{\beta}_\rho^{(j0)}, \boldsymbol{\beta}_\rho^{(j1)}, \ldots, \boldsymbol{\beta}_\rho^{(jp)})$ é o conjunto de coeficientes de regressão associado ao $j-$ésimo intervalo.

Dessa forma, a modelagem é composta pelas seguintes componentes:

1. Equação de observação:

$$T_i \sim MEP(\boldsymbol{\lambda}_\rho^{(i)}, \rho),\ \boldsymbol{\lambda}_\rho^{(i)} = \left(\lambda_\rho^{(i1)}, \lambda_\rho^{(i2)}, \ldots, \lambda_\rho^{(ib)}\right),\ i = 1, \ldots, n.$$

2. Equação de evolução:

$$\boldsymbol{\beta}_\rho^{(j)} = \mathbf{G}_\rho^{(j)} \boldsymbol{\beta}_\rho^{(j-1)} + \mathbf{w}_\rho^{(j)},\ \mathbf{w}_\rho^{(j)} \sim [\mathbf{0}, \mathbf{W}_\rho^{(j)}], \tag{4.7}$$

em que $\mathbf{G}_\rho^{(j)}$ é a matriz conhecida, relacionada ao sistema de evolução, e $\mathbf{w}_\rho^{(j)}$ é o erro acumulado até o intervalo $I_\rho^{(j)}$. A notação $\mathbf{w}_\rho^{(j)} \sim [\mathbf{0}, \mathbf{W}_\rho^{(j)}]$ significa que a distribuição de $\mathbf{w}_\rho^{(j)}$ é parcialmente especificada em termos do vetor de médias e matriz de covariâncias. Assim, a distribuição de $\boldsymbol{\beta}_\rho^{(j)}$ também será parcialmente especificada em termos de seus momentos.

A informação dos dados é processada de maneira sequencial, levando em consideração toda a informação disponível em cada intervalo $I_\rho^{(j)}$ induzido pela partição aleatória ρ. As funções de sobrevivência e densidade de T, dado $T \geq s_{j-1}$, que são necessárias para a análise sequencial, são dadas por

$$S(t \mid T \geq s_{j-1}) = \exp\{-\lambda_\rho^j (t - s_{j-1})\} \tag{4.8}$$

e

$$f(t \mid T \geq s_{j-1}) = \lambda_\rho^j \exp\{-\lambda_\rho^j (t - s_{j-1})\}, \tag{4.9}$$

para $t \in I_\rho^{(j)}, j = 1, \ldots, b$.

Utilizando as funções (4.8) e (4.9), a função de verossimilhança é definida como

$$L = \prod_{j=1}^{b} L_j, \tag{4.10}$$

com

$$L_j = \prod_{i=1}^{n_j} (\lambda_\rho^{(ij)})^{\delta_{ij}} \exp\left\{-\lambda_\rho^{(ij)}(y_{ij} - s_{j-1})\right\},$$

em que n_j é o número de indivíduos que estão sob risco no início de cada intervalo $I_\rho^{(j)}$, e y_{ij} é dado por (3.2). Observe que a função de verossimilhança em (4.10) depende dos coeficientes de regressão pela seguinte relação:

$$\lambda_\rho^{(ij)} = \exp\{\boldsymbol{x}_i^\top \boldsymbol{\beta}_\rho^{(j)}\},$$

com $i = 1, \ldots, n$ e $j = 1, \ldots, b$.

Análise sequencial

A inferência sobre os parâmetros do modelo envolve duas etapas. Na primeira é realizada uma atualização *online*, ou seja, a distribuição *a posteriori* de $\boldsymbol{\beta}_\rho^{(j)}$ é obtida com base em toda informação disponível até o intervalo j, que será denotada aqui por $D_c^{(j)}$, compreendendo a informação dos dados observados e não observados, e a informação subjetiva obtendo-se dessa maneira a distribuição *online* de $(\boldsymbol{\beta}_\rho^{(j)} \mid D_c^{(j)})$. É importante notar que em Gamerman (1991) e Demarqui et al. (2012), $D_c^{(j)}$ é representado somente pelos dados observados. Para a segunda etapa, a distribuição *a posteriori* de $\boldsymbol{\beta}_\rho^{(j)}$ é obtida condicionando-se em toda a informação processada $D_c^{(b)}$, obtendo assim a distribuição suavizada $(\boldsymbol{\beta}_\rho^{(j)} \mid D_c^{(b)})$, $j = 1, \ldots, b$.

Distribuições *online*

A abordagem dinâmica aqui utilizada é formalizada assim como em Gamerman (1991) e Demarqui et al. (2012), em que em cada passo da análise sequencial as distribuições *a priori* e *a posteriori* para os coeficientes da regressão são parcialmente especificadas em termos de seus momentos.

Assuma que a distribuição *a posteriori* para $\boldsymbol{\beta}_\rho^{(j-1)}$ tem vetor de médias $E(\boldsymbol{\beta}_\rho^{(j-1)} \mid D_\rho^{(j-1)}) = \mathbf{m}_{j-1}$ e $Var(\boldsymbol{\beta}_\rho^{(j-1)} \mid D_\rho^{(j-1)}) = \mathbf{C}_{j-1}$, ou seja,

$$\left[\boldsymbol{\beta}_\rho^{(j-1)} \mid D_\rho^{(j-1)}\right] \sim \left[\mathbf{m}_{j-1}, \mathbf{C}_{j-1}\right].$$

Utilizando a equação de evolução em (4.7), temos que a distribuição *a priori* $p(\boldsymbol{\beta}_\rho^{(j)} \mid D_\rho^{(j-1)})$ é dada por

$$\left[\boldsymbol{\beta}_\rho^{(j)} | D_\rho^{(j-1)}\right] \sim \left[\mathbf{a}_j; \mathbf{P}_j\right],$$

em que $\mathbf{a}_j = \mathbf{G}_j \mathbf{m}_{j-1}$ e $\mathbf{P}_j = \mathbf{G}_j \mathbf{C}_{j-1} \mathbf{G'}_j + \mathbf{W}_j$.

A escolha dos parâmetros de suavização pertencentes a diagonal principal da matriz $\mathbf{W}_j$ é uma tarefa complicada, uma vez que temos que especificar quais os valores referentes a variabilidade em cada intervalo. Uma maneira de se contornar essa dificuldade consiste em reescrever $\mathbf{W}_j$ em função de um parâmetro ϕ, denominado fator de desconto, e que permite a utilização de um mecanismo automático para a especificação de $\mathbf{W}_j$. Com esse propósito, re-escrevemos $\mathbf{W}_j$ em termos do fator de desconto $\phi \in (0, 1]$ que controla a quantidade de informação que passa de intervalo para intervalo. Quanto mais próximo de 1 for o valor de ϕ, mais informação passa entre os sucessivos intervalos. Por outro lado, quando $\phi \to 0$, nenhuma informação passa para o intervalo seguinte. Assim, assumindo que $\mathbf{P}_j = (1/\phi)\mathbf{G}_j\mathbf{C}_{j-1}\mathbf{G'}_j$, temos as seguintes relações,

$$\mathbf{P}_j = \mathbf{G}_j\mathbf{C}_{j-1}\mathbf{G'}_j + \mathbf{W}_j \quad \Rightarrow \quad (1/\phi)\mathbf{G}_j\mathbf{C}_{j-1}\mathbf{G'}_j = \mathbf{G}_j\mathbf{C}_{j-1}\mathbf{G'}_j + \mathbf{W}_j.$$

Logo, isolando $\mathbf{W}_j$ temos que

$$\mathbf{W}_j = ((1/\phi) - 1)\mathbf{G}_j\mathbf{C}_{j-1}\mathbf{G'}_j$$

Os momentos *a posteriori* de $[\boldsymbol{\beta}_\rho^{(j)} \mid D_c^{(j)}]$ podem ser obtidos processando-se a informação dos indivíduos sob risco em cada intervalo j de forma sequencial. Como citado anteriormente, assuma que existam n_j indivíduos sob risco no início do $j-$ésimo intervalo e denote por $D_c^{(i-1;j-1)}$, para $i = 1, 2, \ldots, n_j$, toda a informação disponível até $I_\rho^{(j-1)}$, juntamente com a informação processada associada aos primeiros $i-1$ indivíduos que ainda estão sob risco no início do intervalo $I_\rho^{(j)}$. A distribuição condicional *a priori* de $\boldsymbol{\beta}_\rho^{(j)}$, antes de processar a informação do $i-$ésimo indivíduo, é denotada por $[\boldsymbol{\beta}_\rho^{(j)}|D_c^{(i-1;j-1)}] \sim [\mathbf{a}_{ij}; \mathbf{P}_{ij}]$. Uma vez que $\log\left(\lambda_\rho^{(ij)}\right) = \boldsymbol{x}_i^\top\boldsymbol{\beta}_\rho^{(j)}$, a distribuição conjunta *a priori* de $[\boldsymbol{\beta}_\rho^{(j)}, \log\lambda_\rho^{(ij)}]$ é dada por

$$\left[\begin{pmatrix} \boldsymbol{\beta}_\rho^{(j)} \\ \\ \log\lambda_\rho^{(ij)} \end{pmatrix} \mid D_c^{(i-1;j-1)}\right] \sim \left[\begin{pmatrix} \mathbf{a}_{ij} \\ \\ f_{ij} \end{pmatrix}, \begin{pmatrix} \mathbf{P}_{ij} & \mathbf{v}_{ij} \\ \\ \mathbf{v}_{ij}^\top & q_{ij} \end{pmatrix}\right], \tag{4.11}$$

em que $\mathbf{v}_{ij} = \mathbf{P}_{ij}\boldsymbol{x}_i^\top$, $f_{ij} = \boldsymbol{x}_i^\top\mathbf{a}_{ij}$, $\mathbf{a}_{ij} = \mathbf{P}_{ij}\boldsymbol{x}_i$, $q_{ij} = \boldsymbol{x}_i^\top\mathbf{v}_{ij}$ e $\lambda_\rho^{(ij)}$ é a função de taxa de falha para o $i-$ésimo indivíduo no $j-$ésimo intervalo. Para obtermos conjugação, e podermos utilizar a estrutura de agrupamento do MPP, a distribuição gama é assumida como distribuição *a priori* para as taxas de falha $\lambda_\rho^{(ij)}$, ou seja, assumimos que $\lambda_\rho^{(ij)} \sim \text{Gama}(\alpha_{ij}, \gamma_{ij})$. Os hiperparâmetros α_{ij} e γ_{ij} são obtidos através dos momentos de $\lambda_\rho^{(ij)} = \exp\left(\boldsymbol{x}_i^\top\boldsymbol{\beta}_\rho^{(j)}\right)$, que por sua vez são obtidos a partir de (4.11), via aproximação por série de Taylor de primeira ordem, de tal forma que $\alpha_{ij} = q_{ij}^{-1}$ e $\gamma_{ij} = q_{ij}^{-1}\exp(f_{ij})$ (Gamerman, 1994).

Como há conjugação com a componente da função de verossimilhança em (4.10) com a distribuição *a priori* de $\left[\lambda_\rho^{(ij)} \mid D_c^{(i-1;j-1)}\right]$, a distribuição *a posteriori* para $\lambda_\rho^{(ij)}$ é dada por

$$\left[\lambda_\rho^{(ij)} \mid D_c^{(i;j-1)}\right] \sim \text{Gama}\left[\alpha_{ij} + \delta_{ij}, \gamma_{ij} + (y_{ij} - s_{j-1})\right].$$

A distribuição *a posteriori* $[\boldsymbol{\beta}_\rho^{(ij)}|D_c^{(i;j-1)}] \sim [\mathbf{m}_{ij};\mathbf{C}_{ij}]$ é obtida através do método Bayesiano linear descrito em West & Harrison (1997). Assim, atualizamos $[\boldsymbol{\beta}_\rho^{(ij)}|D_c^{i-1;j-1}]$ para a distribuição *a posteriori* $[\boldsymbol{\beta}_\rho^{(ij)}|D_c^{i;j-1}] \sim [\mathbf{m}_{ij};\mathbf{C}_{ij}]$ em que

$$\mathbf{m}_{ij} = \mathbf{a}_{ij} + \frac{\mathbf{v}_{ij}}{q_{ij}} \log\left\{\frac{1+q_{ij}\delta_{ij}}{1+q_{ij}(t_{ij}-a_{j-1})\exp(f_{ij})}\right\}$$

e

$$\mathbf{C}_{ij} = \mathbf{P}_{ij} - \frac{\delta_{ij}}{1+q_{ij}\delta_{ij}}\mathbf{v}_{ij}\mathbf{v}_{ij}^\top.$$

Para cada intervalo tomamos, inicialmente, $\mathbf{a}_{0j} = \mathbf{a}_j$, $\mathbf{P}_{0j} = \mathbf{P}_j$ e $D_c^{(0;j-1)} = D_c^{(j-1)}$. Como a evolução paramétrica em (4.7) é realizada apenas entre os sucessivos intervalos, após o processamento da informação do i-ésimo indivíduo sob risco no intervalo $I_\rho^{(j)}$, tomamos $\mathbf{a}_{i+1;j} = \mathbf{m}_{ij}$ e $\mathbf{P}_{i+1;j} = \mathbf{C}_{ij}$, e repetimos o processo até que toda informação dos n_j elementos seja processada. Em seguida, tomamos $\mathbf{m}_{n_j,j} = \mathbf{m}_j$, $\mathbf{C}_{n_j,j} = \mathbf{C}_j$ e $D_c^{n_j,j} = D_c^{(j)}$. Ao processar toda informação no intervalo j, realizamos a evolução paramétrica através a equação (4.7) e iniciamos o ciclo novamente para os demais intervalos até que toda a informação $D_c^{(b)}$ seja processada.

Distribuição suavizada

Vimos que a distribuição *online*, que a estimativa da taxa de falha de um determinado intervalo é baseada em toda a informação disponível até aquele intervalo. Entretanto, a partir das distribuições *online* obtemos as distribuições *a posteriori* suavizadas para os coeficientes de regressão condicionando na informação disponível em todos os intervalos através de uma análise recursiva.

Segundo Gamerman (1991), a distribuição suavizada para $[\boldsymbol{\beta}_\rho^{(j)}|D_c^{(b)}]$ é dada por

$$[\boldsymbol{\beta}_\rho^{(j)}|D_c^{(b)}] \sim [\mathbf{m}_j^b, \mathbf{C}_j^b], \quad j = 1,\ldots,b,$$

em que

$$\mathbf{m}_j^b = \mathbf{m}_j + \mathbf{C}_j\mathbf{P}_{j+1}^{-1}[\mathbf{m}_{j+1}^b - \mathbf{a}_j] \tag{4.12}$$

e

$$\mathbf{C}_j^b = \mathbf{C}_j - \mathbf{C}_j\mathbf{P}_{j+1}^{-1}[\mathbf{P}_{j+1} - \mathbf{C}_{j+1}^b]\mathbf{P}_{j+1}^{-1}\mathbf{C}_j, \tag{4.13}$$

com $\mathbf{m}_b^b = \mathbf{m}_b$ e $\mathbf{C}_b^b = \mathbf{C}_b$.

Após obter a distribuição suavizada para $[\boldsymbol{\beta}_\rho^{(j)}|D_c^{(b)}]$, podemos obter a distribuição suavizada de $[\lambda_\rho^{(ij)} \mid D_c^{(b)}]$ usando a relação $\lambda_\rho^{(ij)} = \exp(\boldsymbol{x}_i^\top\boldsymbol{\beta}_\rho^{(j)})$, obtendo os momentos de primeira e segunda ordem, de tal forma que

$$\left[\lambda_\rho^{(ij)} \mid D_c^{(b)}\right] \sim \text{Gama}(\alpha_{ij}^b, \gamma_{ij}^b),$$

em que

$$\alpha_{ij}^{b} = \frac{\left[\exp\left(\boldsymbol{x}_i^{\top}\mathbf{m}_j^{b}\right)\right]^2}{\left[\exp\left(\boldsymbol{x}_i^{\top}\mathbf{m}_j^{b}\right)\right]\boldsymbol{x}_i^{\top}\mathbf{C}_j^{b}\boldsymbol{x}\left[\exp\left(\boldsymbol{x}_i^{\top}\mathbf{m}_j^{b}\right)\right]'}$$

e

$$\gamma_{ij}^{b} = \frac{\exp\left(\boldsymbol{x}_i^{\top}\mathbf{m}_j^{b}\right)}{\left[\exp\left(\boldsymbol{x}_i^{\top}\mathbf{m}_j^{b}\right)\right]\boldsymbol{x}_i^{\top}\mathbf{C}_j^{b}\boldsymbol{x}_i\left[\exp\left(\boldsymbol{x}_i^{\top}\mathbf{m}_j^{b}\right)\right]'}.$$

Modelagem da partição aleatória ρ

Seguindo a mesma estratégia descrita no Capítulo 3, para a modelagem da grade τ, temos que a distribuição dos dados completos condicional na partição ρ assume a seguinte forma:

$$\begin{aligned} f(D_c \mid \rho) &= \int_{\boldsymbol{\lambda}_\rho} L(\boldsymbol{\lambda}_\rho, \rho \mid D_c)\, p(\boldsymbol{\lambda}_\rho)\, d\boldsymbol{\lambda}_\rho \\ &= \prod_{j=1}^{b}\prod_{i=1}^{n}\int_{\lambda_\rho^{(ij)}} \left(\lambda_\rho^{(ij)}\right)^{\delta_{ij}} \exp\left\{-\lambda_\rho^{(ij)}(y_{ij}-s_{j-1})\right\} \\ &\quad \times\frac{\gamma_{ij}^{\alpha_{ij}}}{\Gamma(\alpha_{ij})}\left(\lambda_\rho^{(ij)}\right)^{\alpha_{ij}} \exp\left(-\lambda_\rho^{(ij)}\gamma_{ij}\right)\, d\lambda_\rho^{(ij)} \\ &= \prod_{j=1}^{b}\prod_{i=1}^{n}\frac{\gamma_{ij}^{\alpha_{ij}}}{\Gamma(\alpha_{ij})}\frac{\Gamma(\alpha_{ij}+\delta_{ij})}{(\gamma_{ij}+y_{ij}-s_{j-1})^{\alpha_{ij}+\delta_{ij}}}. \end{aligned} \tag{4.14}$$

Assumimos que a distribuição *a priori* para ρ é da mesma forma como em (3.8). Então, a distribuição condicional completa de ρ é dada por

$$\begin{aligned} p(\rho \mid D_c) &= f(D_c \mid \rho)p(\rho) \\ &\propto \prod_{j=1}^{b}\prod_{i=1}^{n}\frac{\gamma_{ij}^{\alpha_{ij}}}{\Gamma(\alpha_{ij})}\frac{\Gamma(\alpha_{ij}+\delta_{ij})}{(\gamma_{ij}+t_{ij}-s_{j-1})^{\alpha_{ij}+\delta_{ij}}}c_{I_\rho^{(j)}}. \end{aligned} \tag{4.15}$$

Utilizamos o algoritmo proposto em Barry & Hartigan (1993) para gerar de $p(\rho \mid D_c)$. Um ponto importante a ressaltar é que, condicional nos dados completos, ou seja, nos pseudo-tempos de falha, a expressão (4.15) é a mesma encontrada em Demarqui et al. (2012).

Uma estimativa *a posteriori* para a função de sobrevivência para um novo indivíduo com vetor de covariáveis $\boldsymbol{x}$ é dada por

$$S(t \mid D_c) = \sum_{\rho} S(t \mid D_c, \rho)p(\rho \mid D_c), \tag{4.16}$$

em que,

$$\begin{aligned} S(t \mid D_c, \rho) &= \int S(t \mid \boldsymbol{\lambda}_\rho^{(\boldsymbol{x})}, \rho)p(\boldsymbol{\lambda}_\rho^{(\boldsymbol{x})})d\boldsymbol{\lambda}_\rho^{(\boldsymbol{x})} \\ &= \left(1+\frac{t-s_{j-1}}{\gamma_j^{(\boldsymbol{x})}}\right)^{\alpha_j^{(\boldsymbol{x})}}\prod_{r=1}^{j-1}\left(1+\frac{s_r-s_{r-1}}{\gamma_r^{(\boldsymbol{x})}}\right)^{\alpha_r^{(\boldsymbol{x})}}, \; t \in I_\rho^{(j)}, \end{aligned} \tag{4.17}$$

e $\boldsymbol{\lambda}_\rho^{(\boldsymbol{x})} = (\boldsymbol{\lambda}_\rho^{(1;\boldsymbol{x})}, \ldots, \boldsymbol{\lambda}_\rho^{(b;\boldsymbol{x})})$. Observe que em (4.16), a função de sobrevivência *a posteriori* é obtida tomando a média sobre todas as partições $\rho = (i_0, i_1, \ldots, i_b)$.

A distribuição *a posteriori* para $\boldsymbol{\beta}_k$, com $k = 1, \ldots, m^{'}$ é determinada como uma mistura de distribuições, ou seja,

$$p(\boldsymbol{\beta}_k \mid D) = \sum_{i_{j-1}<k<i_j} p(\boldsymbol{\beta}_\rho^{(j)} \mid D_c^{(b)}) R(I_\rho^{(j)}),$$

em que $p(\boldsymbol{\beta}_\rho^{(j)} \mid D_c^{(b)})$, é a distribuição *a posteriori* suavizada de $\boldsymbol{\beta}_\rho^{(j)}$, para $j = 1, \ldots, b$.

4.1.4 Amostragem dos pseudotempos de falha

Na presença de dados de sobrevivência com censura intervalar, vimos que os modelos desenvolvidos aqui são condicionais nos pseudotempos de falha. Assim, nesta seção apresentamos os passos para a amostragem dos pseudotempos de falha, assumindo o MEP para a função risco basal.

De forma simplificada, considere $D = \{(l_i, r_i] : i = 1, 2, \ldots, n\}$ como sendo os valores dos intervalos observados. Ao iniciar o algoritmo, tomamos os primeiros pseudotempos de falha como os pontos médios dos intervalos observados e para as censuras como os limites inferiores dos intervalos observados, ou seja,

$$t_i = \begin{cases} \dfrac{l_i + r_i}{2}, & \text{se} \quad \delta_i = 1 \\ l_i, & \text{se} \quad \delta_i = 0 \end{cases},$$

No caso do MEP com efeito das covariáveis fixo no tempo temos a seguinte função de distribuição acumulada,

$$\begin{aligned} F(t \mid \boldsymbol{\lambda}_\rho, \rho, \boldsymbol{\beta}, \boldsymbol{x}, \tau) &= 1 - \exp\left\{-\exp\left(\boldsymbol{x}^\top \boldsymbol{\beta}\right) \lambda_\rho^{(j)}(t - s_{j-1})\right\} \\ &\quad \times \exp\left\{-\exp\left(\boldsymbol{x}^\top \boldsymbol{\beta}\right) \sum_{g=1}^{j-1} \lambda_\rho^{(g)}(s_j - s_{j-1})\right\}, \quad t \in I_j. \end{aligned} \tag{4.18}$$

Assim, para este modelo, e seguindo os passos do algoritmo genérico descrito na Subseção 2.4.3, temos a forma de geração dos pseudotempos de falha são obtidos pela seguinte expressão

$$t_i = \frac{-\log(1 - u_i) - \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \sum_{g=1}^{j-1} \lambda_\rho^{(g)} \Delta_g}{\lambda_\rho^{(j)} \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})} + s_{j-1}, \quad t_i \in I_j, \text{ para } \delta_i = 1, \tag{4.19}$$

em que $\Delta_g = s_g - s_{g-1}$ e u_i é um valor gerado de uma distribuição uniforme no intervalo $[F(l_i); F(r_i)]$ para o $i-$ésimo indivíduo, sendo $F(\cdot)$ a função de distribuição do MEP avaliadas nos limites inferior e superior dos intervalos observados.

Para os modelos com efeito variando no tempo apresentados na Seção 4.1.3, imputamos os pseudotempos de falha para cada intervalo através da expressão

$$t_i = \frac{-\log(1-u_i) - \sum_{g=1}^{j-1} \lambda_\rho^{(ig)} \Delta_g}{\lambda_\rho^{(ij)}} + s_{j-1}, \quad t_i \in I_\rho^{(j)}, \ \forall \delta_i = 1, \tag{4.20}$$

em que $\lambda_\rho^{(ij)} = \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j\right)$ para $t_i \in I_\rho^{(j)}$.

Ilustramos graficamente através da Figura 4.1 o esquema de geração de um pseudotempo de falha. Observe que, gerado o valor de $u_i \sim \text{Unif}[F(l_i); F(r_i)]$ o valor de t_i é obtido diretamente.

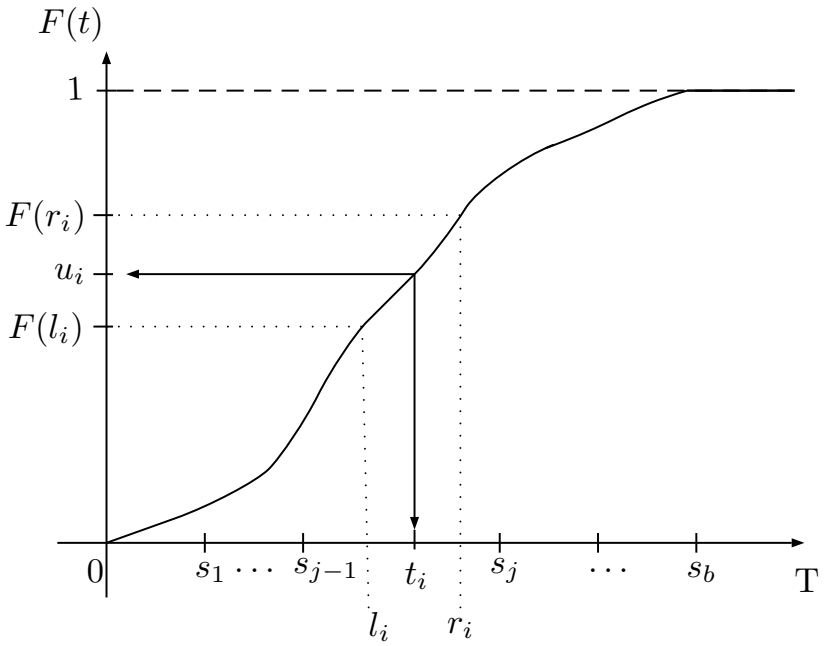

Figura 4.1: Geração dos pseudotempos de falha.

4.2 Modelo com fração de cura

Dados de sobrevivência sujeitos a censura intervalar também podem apresentar uma parcela de indivíduos na população que não são suscetíveis à ocorrência do evento de interesse, caracterizando assim um cenário com fração de curados. Por exemplo, em uma determinada população de fumantes, podemos ter o interesse em estudar o tempo até a interrupção do fumo (Banerjee & Carlin, 2004). Neste caso, uma parcela dos indivíduos continuarão sendo fumantes ativos. Kim et al. (1993), Sun (2006) e Hashimoto et al. (2015) utilizam dados de tempo até a soro conversão de HIV-1 em pacientes hemofílicos. Neste caso, dependendo do tratamento e outras características, indivíduos podem não desenvolver a soroconversão, representando uma população com fração de curados. Observe que em ambos os casos, não

há possibilidade de o tempo de falha ser observado de forma exata, pois não sabemos exatamente quando o indivíduos parou de fumar, assim como, quando houve a soroconversão, caracterizando assim exemplos de dados de sobrevivência sujeitos a censura intervalar.

Nesta seção apresentamos um novo modelo de fração de cura de tempos de promoção para dados de sobrevivência sujeitos a censura intervalar. Utilizamos aqui o algoritmo de ampliação de dados descrito na Subseção 2.4.3. O MEP com grade aleatória é assumido para a modelagem dos tempos de promoção, e uma estrutura dinâmica é considerada para modelar as taxas de falha do MEP.

Seja $T = \min(R_0, R_1, R_2, \ldots, R_M)$ uma variável aleatória, não observável, representando o tempo até a ocorrência do evento de interesse. Assumimos que $R \sim MEP(\boldsymbol{\lambda}_\rho, \tau)$, ρ é a partição aleatória e $\boldsymbol{\lambda}_\rho = (\lambda_\rho^{(1)}, \lambda_\rho^{(2)}, \ldots, \lambda_\rho^{(b)})$ é o vetor de taxas do MEP, e τ é a grade dos tempos induzidos pela partição ρ. Denote por $D_c = \{(y_i, \delta_i, \boldsymbol{z}_i, m_i) : i = 1, \ldots, n\}$ denotando os dados aumentados. Assim a função de verossimilhança em (2.29) assume a seguinte expressão

$$\begin{aligned} L(\boldsymbol{\lambda}_\rho, \boldsymbol{\beta}, \boldsymbol{\psi}|D_c) &= \prod_{i=1}^{n}\left[\prod_{j=1}^{b} m_i^{\delta_{ij}} \left(\lambda_\rho^{(j)}\right)^{\delta_{ij}} \exp\left\{-m_i\lambda_\rho^{(j)}(y_{ij} - s_{j-1})\right\} P(M_i = m_i)\right] \\ &= \prod_{j=1}^{b} \left(\lambda_\rho^{(j)}\right)^{\nu_j} \exp\left\{-\lambda_\rho^{(j)}\xi_j\right\} \\ &\quad \times \exp\left\{\sum_{i=1}^{n}\left(\delta_{ij}\log(m_i) + m_i\log(\theta_i) - \log(m_i) - \theta_i\right)\right\} \end{aligned} \tag{4.21}$$

em que $\xi_j = \sum_{i=1}^{n} m_i(y_{ij} - s_{j-1})$, $j = 1, ..., b$ e $\theta_i = \exp\left\{\boldsymbol{z}_i^\top\boldsymbol{\psi}\right\}$. Note que em D_c, tanto y_i (para $\delta_i = 1$) quanto m_i são valores não observáveis. Além disso, assim como em Demarqui et al. (2014), a função de verossimilhança dada em (4.21) (como função das taxas) fatora em um produto de núcleos de distribuições gama. Esse fato permite utilizar a estrutura de agrupamento do MPP para modelar a grade do MEP.

4.2.1 Modelagem dinâmica

Um caso particular do modelo dinâmico proposto em Demarqui et al. (2012) é o modelo desenvolvido por Demarqui (2010), que por sua vez corresponde a uma extensão do modelo dinâmico desenvolvido por Gamerman (1994) ao considerar a grade τ como sendo uma quantidade aleatória. Assim, apresentamos a seguir um novo modelo de fração de cura dinâmico baseado na estrutura do modelo de tempos de promoção, assumindo o MEP com grade aleatória (Demarqui et al., 2008) e correlacionando as taxas do MEP (Gamerman, 1994), para dados de sobrevivência com censura intervalar.

O modelo com fração de cura dinâmico é determinado da seguinte forma:

i) $R \sim MEP(\boldsymbol{\lambda}_\rho, \rho)$, $\boldsymbol{\lambda}_\rho = (\lambda_\rho^{(1)}, \lambda_\rho^{(2)}, \ldots, \lambda_\rho^{(b)})$.

ii) $\left[\lambda_\rho^{(j)} \mid D_c^{(j-1)}\right] \sim \text{Gama}(\phi\alpha_{j-1}; \phi\gamma_{j-1}), \quad j = 1, \ldots, b.$

em que $D_c^{(j-1)}$ representa toda informação disponível até o intervalo $j-1$, ou seja, a informação dos dados observados e não observados e a informação subjetiva, α_{j-1} e γ_{j-1} correspondem aos parâmetros da distribuição a posteriori de $lambda_{rho}^{(j-1)}$, e ϕ representa o fator de desconto, e controla a quantidade de informação que é passada entre os sucessivos intervalos. Diferentemente do que foi proposto em Gamerman (1994) e Demarqui (2010), introduzimos uma distribuição *a priori* para modelar o fator de desconto ϕ, o que é uma vantagem pois evitamos uma análise de sensibilidade para o mesmo. No item *ii*), do modelo com fração de cura dinâmico definido acima, assumimos uma distribuição gama para garantir a conjugação e assim utilizar a estrutura de agrupamento do MPP para a modelagem da grade do MEP.

Seja $D_c^{(0)}$ a informação inicial, totalmente subjetiva, disponível antes de os dados serem obtidos. Sendo assim, os passos para a análise sequencial são os seguintes:

1. Tomar $j = 1$ e especificar $\left[\lambda_\rho^{(j)} \mid D_c^{(j-1)}, \rho\right] \sim \text{Gama}(\alpha_{j-1}; \gamma_{j-1})$;
2. Atualizar a informação *a priori* contida em $\left[\lambda_\rho^{(j)} \mid D_c^{(j-1)}, \rho, \phi\right]$ com a informação dos dados disponível em $I_\rho^{(j)}$, resultando em $\left[\lambda_\rho^{(j)} \mid D_c^{(j)}, \rho, \phi\right] \sim \text{Gama}(\alpha_j; \gamma_j)$, em que $\alpha_j = \alpha_{j-1} + \nu_j$ e $\gamma_j = \gamma_{j-1} + \xi_j$, com $\nu_j = \sum_{i=1}^n \delta_{ij}$, $\xi_j = \sum_{i=1}^n m_i(y_{ij} - s_{j-1})$ e y_{ij} como definido em (3.2);
3. Realizar a evolução paramétrica: $\left[\lambda_\rho^{(j+1)} \mid D_c^{(j)}, \rho, \phi\right] \sim \text{Gama}\left(\phi\alpha_j \; ; \; \phi\gamma_j\right)$;
4. Fazer $j = j + 1$, retornar ao passo (2) e repetir o ciclo até que toda informação seja processada.

De maneira similar ao desenvolvimento do modelo descrito na Subseção 4.1.3, a análise sequencial acima está relacionada com a seguinte equação de evolução:

$$\log\left(\lambda_\rho^{(j)}\right) = \log\left(\lambda_\rho^{(j-1)}\right) + w_j, \quad w_j \sim [0, W_j], \quad j = 1, \ldots, b, \tag{4.22}$$

em que $w_j \sim [0, W_j]$ corresponde à distribuição dos erros, que é parcialmente especificada em termos de $E(w_j) = 0$ e $Var(w_j) = W_j$ em que

$$W_j = \left(\frac{1}{\phi} - 1\right) Var\left[\log\left(\lambda_\rho^{(j-1)}\right) | D_c^{(j-1)}\right], \tag{4.23}$$

para $j = 1, \ldots, b$.

A dependência entre os elementos do vetor $\boldsymbol{\lambda}_\rho$ associados aos intervalos induzidos pela partição aleatória ρ é introduzida via hiperparâmetros da distribuição *a priori*. Dessa forma, os componentes do vetor $\boldsymbol{\lambda}_\rho$ são condicionalmente independentes e, consequentemente, a distribuição conjunta *a priori* para $(\boldsymbol{\lambda}_\rho, \rho)$ pode ser expressa na forma produto, o que garante

que a abordagem proposta por Demarqui et al. (2008) possa ser utilizada para a modelagem da grade τ. Além disso, a introdução da dependência das taxas de falha através dos hiperparâmetros da distribuição gama nos permite a estimação do fator de desconto, diferentemente do modelo descrito na Subseção 4.1.3.

Dessa forma, a distribuição *a priori* conjunta para $(\boldsymbol{\lambda}_\rho, \boldsymbol{\psi}, \phi, \rho)$ é dada por

$$p(\boldsymbol{\lambda}_\rho, \phi, \boldsymbol{\psi}, \rho) = p(\boldsymbol{\lambda}_\rho \mid \phi, \rho)p(\phi)p(\rho)p(\boldsymbol{\psi}). \tag{4.24}$$

Assumimos para a partição aleatória ρ a seguinte distribuição *a priori*

$$p(\rho) \propto \prod_{j=1}^{b} c_{I_\rho^{(j)}}. \tag{4.25}$$

Para o fator de desconto ϕ, atribuímos uma distribuição *a priori* beta com parâmetros θ_1 e θ_2, cuja função densidade é

$$p(\phi) = \frac{\Gamma(\theta_1 + \theta_2)}{\Gamma(\theta_1)\Gamma(\theta_2)}\phi^{\theta_1 - 1}(1-\phi)^{\theta_2 - 1}. \tag{4.26}$$

em que $\theta_1 > 0$ e $\theta_2 > 0$.

Condicional à partição aleatória ρ e no fator de desconto ϕ, a distribuição *a priori* de $[\boldsymbol{\lambda}_\rho \mid \phi, \rho]$ é expressa da seguinte forma:

$$p(\boldsymbol{\lambda}_\rho \mid \phi, \rho) = \frac{\gamma_0^{\alpha_0}}{\Gamma(\alpha_0)} \left(\lambda_\rho^{(1)}\right)^{\alpha_0 - 1} e^{\lambda_\rho^{(1)}\gamma_0} \prod_{j=2}^{b} \frac{(\phi\gamma_{j-1})^{\phi\alpha_{j-1}}}{\Gamma(\phi\alpha_{j-1})} \left(\lambda_\rho^{(j)}\right)^{\phi\alpha_{j-1} - 1} e^{\lambda_\rho^{(j)}\phi\gamma_{j-1}}. \tag{4.27}$$

Finalmente, assumimos uma distribuição normal para cada componente do vetor $\boldsymbol{\psi}$ como a distribuição *a priori*. Então, substituindo os elementos na distribuição *a priori* conjunta de $[\boldsymbol{\lambda}_\rho,\ \phi\ \boldsymbol{\psi},\ \rho]$ em (4.24) temos

$$\begin{aligned} p(\boldsymbol{\lambda}_\rho, \phi, \boldsymbol{\psi}, \rho) \quad \propto \quad & \frac{\gamma_0^{\alpha_0}}{\Gamma(\alpha_0)} \left(\lambda_\rho^{(1)}\right)^{\alpha_0 - 1} e^{\lambda_\rho^{(1)}\gamma_0} c_{I_\rho^{(1)}} \\ & \times \prod_{j=2}^{b} \frac{(\phi\gamma_{j-1})^{\phi\alpha_{j-1}}}{\Gamma(\phi\alpha_{j-1})} \left(\lambda_\rho^{(j)}\right)^{\phi\alpha_{j-1} - 1} e^{\lambda_\rho^{(j)}\phi\gamma_{j-1}} c_{I_\rho^{(j)}} \\ & \times \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\psi} - \mu)^2\right\}. \end{aligned} \tag{4.28}$$

Assim, a distribuição *a posteriori* para o modelo de tempo de promoção dinâmico é dada

por

$$\begin{aligned} p(\boldsymbol{\lambda}_\rho, \boldsymbol{\psi}, \rho, \phi \mid D_c) &\propto L(\boldsymbol{\lambda}_\rho, \boldsymbol{\psi} \mid D_c) p(\boldsymbol{\lambda}_\rho \mid \rho, \phi) p(\rho) p(\phi) p(\boldsymbol{\psi}) \\ &\propto \left(\lambda_\rho^{(1)}\right)^{\nu_1+\alpha_0} \exp\left\{-\lambda_\rho^{(1)}(\xi_1+\gamma_0)\right\} c_{I_\rho^{(1)}} \\ &\quad \times \prod_{j=2}^{b} \left(\lambda_\rho^{(j)}\right)^{\nu_j+\phi\alpha_{j-1}-1} \exp\left\{-\lambda_\rho^{(j)}(\xi_j+\phi\gamma_{j-1})\right\} c_{I_\rho^{(j)}} \\ &\quad \times \exp\left\{\sum_{i=1}^{n} m_i \boldsymbol{z}_i^\top \boldsymbol{\psi} - \log(m_i!) - e^{\boldsymbol{z}_i^\top \boldsymbol{\psi}}\right\} \\ &\quad \times \phi^{\theta_1-1}(1-\phi)^{\theta_2-1} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\psi}-\mu)^2\right\}. \end{aligned} \tag{4.29}$$

Devido ao fato de que a distribuição *a posteriori* em (4.29) não apresentar forma analítica conhecida, métodos MCMC serão empregados para a realização da análise bayesiana. Especificamente, com o intuito de melhorar a convergência da cadeia, será utilizado o algoritmo de Gibbs colapsado proposto por Liu (1994).

A distribuição condicional completa para $\boldsymbol{\psi}$ tem a seguinte forma:

$$\begin{aligned} p(\boldsymbol{\psi} \mid \boldsymbol{\lambda}_\rho, \rho, D_c^*) &\propto L(\boldsymbol{\lambda}_\rho, \boldsymbol{\psi} \mid D_c^*) p(\boldsymbol{\psi}) \\ &\propto \exp\left\{\sum_{i=1}^{n} \delta_i \boldsymbol{z}_i^\top \boldsymbol{\psi} - \exp(\boldsymbol{z}_i^\top \boldsymbol{\psi}) \exp\left(-\boldsymbol{\lambda}_\rho (y_{ij} - s_{j-1})\right)\right\} p(\boldsymbol{\psi}), \end{aligned}$$

em que $D_c^* = \{(y_i, \delta_i, \boldsymbol{z}_i) : i = 1, \ldots, n\}$, contém apenas os pseudotempos de falha como quantidades latentes, e $L(\boldsymbol{\lambda}_\rho, \boldsymbol{\psi} \mid D_c^*)$ é a função de verossimilhança obtida pela soma em M, como em (2.20). Pode ser mostrado que a distribuição condicional completa de $[\boldsymbol{\psi} \mid \boldsymbol{\lambda}_\rho, \rho, D_c]$ é log-côncava para cada componente de $\boldsymbol{\psi}$. Dessa forma, o algoritmo ARS pode ser utilizado para amostrar de $[\boldsymbol{\psi} \mid \boldsymbol{\lambda}_\rho, \rho, D_c]$.

Conforme discutido na Subseção 2.3.1, a distribuição condicional para o número de causas latentes é

$$[M_i \mid \boldsymbol{\psi}, \boldsymbol{\lambda}_\rho, \rho, D_c^*] \sim \text{Poisson}\left(\exp\left(\boldsymbol{z}_i^\top \boldsymbol{\psi} - \sum_{j=1}^{b} \lambda_\rho^{(j)}(y_{ij} - s_{j-1})\right)\right) + \delta_i, \tag{4.30}$$

para $i = 1, \ldots, n$. Logo, amostras de $[M_i \mid \boldsymbol{\psi}, \boldsymbol{\lambda}_\rho, \rho, D_c^*]$ podem ser obtidas de maneira direta.

Baseando-se na função de verossimilhança em (4.21) e na distribuição *a priori* em (4.27),

a distribuição dos dados completos, condicional em $\boldsymbol{\psi}$, ϕ e ρ, tem a seguinte forma:

$$\begin{aligned} f(D_c \mid \phi, \boldsymbol{\psi}, \rho) &\propto \prod_{j=1}^{b} \int_{\lambda_\rho^{(j)}} L(\Theta, \boldsymbol{\psi} \mid D_c) p(\lambda_\rho^{(j)}) d\lambda_\rho^{(j)} \\ &\propto \frac{\gamma_0^{\alpha_0}}{\Gamma(\alpha_0)} \frac{\Gamma(\nu_1 + \alpha_0)}{(\xi_1 + \gamma_0)^{\nu_1 + \alpha_0}} \\ &\quad \times \prod_{j=2}^{b} \frac{(\phi \gamma_{j-1}^{\phi \alpha_{j-1}})}{\Gamma(\phi \alpha_{j-1})} \frac{\Gamma(\nu_j + \phi \alpha_{j-1})}{(\xi_j + \phi \gamma_{j-1})^{\nu_j + \phi \alpha_{j-1}}}. \end{aligned} \tag{4.31}$$

Dessa forma, a distribuição condicional completa para ρ é dada por

$$\begin{aligned} p(\rho \mid D_c, \phi) &\propto \prod_{j=1}^{b} \int L(\Theta, \boldsymbol{\psi} \mid D_c) p(\lambda_\rho^{(j)}) d\lambda_\rho^{(j)} \\ &\propto \frac{\gamma_0^{\alpha_0}}{\Gamma(\alpha_0)} \frac{\Gamma(\nu_1 + \alpha_0)}{(\xi_1 + \gamma_0)^{\nu_1 + \alpha_0}} c_{I_\rho^{(1)}} \\ &\quad \times \prod_{j=2}^{b} \frac{(\phi \gamma_{j-1}^{\phi \alpha_{j-1}})}{\Gamma(\phi \alpha_{j-1})} \frac{\Gamma(\nu_j + \phi \alpha_{j-1})}{(\xi_j + \phi \gamma_{j-1})^{\nu_j + \phi \alpha_{j-1}}} c_{I_\rho^{(j)}}. \end{aligned} \tag{4.32}$$

Utilizamos o algoritmo de Barry & Hartigan (1993) para realizar a atualização da partição aleatória ρ. Pela equação (4.31), temos que a distribuição condicional completa para ϕ tem a seguinte expressão:

$$p(\phi \mid \rho, D_c) \propto \prod_{j=2}^{b} \frac{(\phi \gamma_{j-1})^{\phi \alpha_{j-1}}}{\Gamma(\phi \alpha_{j-1})} \frac{\Gamma(\nu_j + \phi \alpha_{j-1})}{(\xi_j + \phi \gamma_{j-1})^{\nu_j + \phi \alpha_{j-1}}} \phi^{\theta_1 - 1} (1 - \phi)^{\theta_2 - 1}. \tag{4.33}$$

Como a distribuição $[\phi|\rho, D_c]$ não é log-côncava, utilizamos o algoritmo da amostragem por rejeição adaptativa com passo de metropolis (ARMS), proposto por Gilks et al. (1995), para obter amostras de $[\phi|\rho, D_c]$.

A distribuição condicional completa para o vetor $\boldsymbol{\lambda}_\rho$ é dada por

$$\begin{aligned} p(\boldsymbol{\lambda}_\rho \mid \boldsymbol{\psi}, \rho, \phi, D_c) &\propto L(\Theta, \boldsymbol{\psi} \mid D_c) p(\boldsymbol{\lambda}_\rho \mid \rho, \phi) \\ &\propto \left[\left(\lambda_\rho^{(1)}\right)^{\nu_1 + \alpha_0} \exp\left\{ -\lambda_\rho^{(1)} (\xi_1 + \gamma_0) \right\} \right] \\ &\quad \times \left[\prod_{j=2}^{b} \left(\lambda_\rho^{(j)}\right)^{\nu_j + \phi \alpha_{j-1} - 1} \exp\left\{ -\lambda_\rho^{(j)} (\xi_j + \phi \gamma_{j-1}) \right\} \right]. \end{aligned} \tag{4.34}$$

A distribuição condicional completa acima é expressa em termos das distribuições *online* para o vetor $\boldsymbol{\lambda}_\rho$. No entanto, inferências sobre as componentes do vetor $\boldsymbol{\lambda}_\rho$ são realizadas com base nas distribuições *a posteriori* suavizadas, que podem ser obtidas utilizando-se o algoritmo recursivo apresentado na Subseção 4.1.3, considerando-se o vetor $\boldsymbol{\beta}_\rho^{(j)}$ como um vetor contendo apenas o intercepto.

Como consequência, a distribuição condicional completa para λ_k, com $k = 1, \ldots, m'$, ou seja,

as taxas de falha associadas à partição inicial, é dada por

$$p(\lambda_k \mid \boldsymbol{\psi}, \phi, \rho, D_c) = \sum_{i_{j-1}<k\leq j} p(\lambda_\rho^{(j)} \mid \boldsymbol{\psi}, \phi, \rho, D_c^{(b)}) R(I_\rho^{(j)} \mid \boldsymbol{\psi}, \phi, \rho, D_c^{(b)}),$$

em que a quantidade $R(I_\rho^{(j)} \mid \boldsymbol{\psi}, \phi, \rho, D_c^{(b)})$ é chamada de relevância *a posteriori* e representa a probabilidade de cada intervalo $I_\rho^{(j)}$ pertencer à partição aleatória ρ, e $p(\lambda_\rho^{(j)} \mid \boldsymbol{\psi}, \phi, \rho, D_c^{(b)})$ representa a distribuição *a posteriori* suavizada para as taxas comuns $\lambda_\rho^{(j)}$, para $j = 1, \ldots, b$ e $b \in \{1, \ldots, m^{'}\}$.

Seguindo os passos descritos na Subseção 4.1.4, para a geração dos pseudotempo de falha utilizamos a função de distribuição associada aos indivíduos não curados, uma vez que os elementos que falharam são os não curados. Dessa forma, pela expressão (2.14), temos que a função de distribuição para os indivíduos não curados é dada por

$$F_{NC}(t \mid \boldsymbol{\lambda}_\rho, \rho, \boldsymbol{z}, \boldsymbol{\psi}, \tau) = 1 - \frac{e^{-\theta F(t)} - e^{-\theta}}{1 - e^{-\theta}},$$

em que $F(t)$ é a função de distribuição acumulada do MEP e $\theta = \exp\left(\boldsymbol{z}^\top \boldsymbol{\psi}\right)$. Então, temos que a expressão utilizada para a geração dos pseudotempos de falha é dada por

$$t_i = -\frac{1}{\lambda_\rho^{(j)}} \left\{ \log\left[1 + \frac{\log((1-u_i)(1-\exp(\theta_i)) + \exp(\theta_i))}{\theta_i}\right] + \sum_{g=1}^{j-1} \lambda_\rho^{(g)}(s_j - s_{j-1}) \right\} + s_{j-1}, \tag{4.35}$$

para $t_i \in I_\rho^{(j)}$ e $i = 1 \ldots, n$. O valor u_i é gerado por uma distribuição uniforme no intervalo $[F_{NC}(l_i); F_{NC}(r_i)]$ para o $i-$ésimo indivíduo.

4.3 Seleção de modelos

Com o objetivo de comparar os modelos descritos nas seções anteriores, serão utilizados alguns critérios bayesianos para comparação de modelos, a saber: o logaritmo da pseudoverossimilhança marginal ($LPML$), critério de informação da desviância (DIC) e critério de informação de Watanabe ($WAIC$).

4.3.1 Ordenada da preditiva condicional

A ordenada preditiva condicional (CPO) é uma abordagem de validação cruzada bastante utilizada na literatura Bayesiana (Gelfand et al., 1992). Segundo Sinha et al. (1999), dado um modelo de sobrevivência para dados sujeitos a censura intervalar, a estatística CPO para o $i-$ésimo individuo é definida como

$$CPO_i = P(T_i \in (l_i, r_i] | D^{(-i)}),$$

em que $D^{(-i)}$ são os dados observados sem a $i-$ésima observação. Essa estatística corresponde à probabilidade preditiva *a posteriori* para a $i-$ésima observação condicional em todas as outras observações no modelo. Valores grandes de CPO_i indicam que a $i-$ésima observação favorece o modelo ajustado. De acordo com Sinha et al. (1999), a CPO_i pode ser calculada utilizando a seguinte expressão

$$CPO_i = \left(E\left[\frac{1}{P(T_i \in (l_i, r_i]|\boldsymbol{\Theta})} \right] \mid D \right)^{-1}, \tag{4.36}$$

com $P(T_i \in (l_i, r_i]|\boldsymbol{\Theta})$ sendo a contribuição do $i-$ésimo indivíduo para a função de verossimilhança.

A esperança em (4.36) é tomada com respeito à distribuição *a posteriori* conjunta de $\boldsymbol{\Theta}$ dado D. Como discutido em Dey et al. (1997), a CPO_i pode ser calculada através de uma média harmônica das cópias de $P(T_i \in (l_i, r_i]|\boldsymbol{\Theta})$ avaliadas pelas amostras MCMC da distribuição *a posteriori* de $\boldsymbol{\Theta}$. Então, seja $\Theta^{(1)}, \Theta^{(2)}, \ldots, \Theta^{(Q)}$ uma amostra de tamanho Q da distribuição *a posteriori* $p(\boldsymbol{\Theta}|D)$. Assim, a aproximação de Monte Carlo de CPO_i é dada por

$$\widehat{CPO_i} = \left(\frac{1}{Q} \sum_{q=1}^{Q} \left[\frac{1}{P(T_i \in (l_i, r_i]|\boldsymbol{\Theta}^{(q)})} \right] \right)^{-1}, i = 1, \ldots, n. \tag{4.37}$$

Uma quantidade bastante utilizada para a comparação de modelos Bayesianos pode ser obtida resumindo os valores das $CPO_i's$. Ela é denominada $LPML$, e pode ser estimada da seguinte forma

$$LPML = \sum_{i=1}^{n} \log(\widehat{CPO_i}).$$

O modelo com maior valor de $LPML$ é melhor em termos de ajuste.

4.3.2 Critério de informação da desviância

O critério de informação da desviância (DIC) proposto por Spiegelhalter et al. (2002) é outra medida de qualidade de ajuste e complexidade do modelo. A desviância (*deviance*) possui a seguinte expressão

$$D(\boldsymbol{\Theta}) = -2 \sum_{i=1}^{n} \log \left[P(T_i \in (l_i, r_i]|\boldsymbol{\Theta}) \right],$$

em que $P(T_i \in (l_i, r_i]|\boldsymbol{\Theta})$ é a contribuição na verossimilhança referente a $i-$ésima observação. Sejam $\bar{\boldsymbol{\Theta}}$ e $\bar{D} = E(D(\boldsymbol{\Theta}) \mid D)$ a média *a posteriori* para $\boldsymbol{\Theta}$ e $D(\boldsymbol{\Theta})$, respectivamente. Dessa forma, o DIC é definido por

$$DIC = \bar{D} - 2\, p_D,$$

em que $p_D = D(\bar{D})$ é o número efetivo de parâmetros.

A média *a posteriori* da desviância pode ser aproximada através da amostra MCMC por

$\bar{D} = \sum_{q=1}^{Q} D(\boldsymbol{\Theta}_q)/Q$. Gelman et al. (2003) nos mostram que o DIC pode ser estimado por

$$\widehat{DIC} = \bar{D} + 0.5\ p_D,$$

em que p_D pode ser estimado utilizando a seguinte expressão:

$$\widehat{p_D} = \widehat{Var}(D(\boldsymbol{\Theta})) = \frac{1}{Q-1}\sum_{q=1}^{Q}[D(\Theta^{(q)}) - \bar{D}]^2.$$

Sob este critério, o modelo que melhor se adequa aos dados é aquele que apresentar o DIC de menor valor.

4.3.3 Critério de informação de Watanabe

O critério de informação de Watanabe, proposto por Watanabe (2010) é mais um critério de seleção de modelos Bayesianos que tem sido bastante utilizado na literatura, e que pode ser visto como o critério DIC aprimorado (Gelman et al., 2013). O logaritmo da densidade preditiva é definido como

$$\text{lpd} = \sum_{i=1}^{n} \log\left(p_{post}(y_i)\right) = \sum_{i=1}^{n} \log\left(\int P(T_i \in (l_i, r_i]|\boldsymbol{\Theta}) p_{post}(\boldsymbol{\Theta}) d\boldsymbol{\Theta}\right),$$

em que $p_{post}(y_i)$ é a distribuição preditiva *a posteriori*. Na prática, o lpd pode ser aproximado utilizando as amostras geradas pela distribuição *a posteriori* $p(\boldsymbol{\Theta} \mid D)$. Dessa forma,

$$\widehat{\text{lpd}} = \sum_{i=1}^{n} \log\left(\frac{1}{Q}\sum_{q=1}^{Q} P(T_i \in (l_i, r_i]|\boldsymbol{\Theta}^s)\right),$$

em que $P(T_i \in (l_i, r_i]|\boldsymbol{\Theta}^s)$ representa a função de densidade dos dados observados avaliada na $s-$ésima amostra *a posteriori* de $\boldsymbol{\Theta}$. Definimos a variância do logaritmo da densidade preditiva como

$$\text{pd} = \sum_{i=1}^{n} Var_{post}\left[\log\left(P(T_i \in (l_i, r_i]|\boldsymbol{\Theta})\right)\right],$$

em que a aproximação pode ser computada da seguinte forma:

$$\widehat{\text{pd}} = \sum_{i=1}^{n} V_{s=1}^{S}\left[\log\left(P(T_i \in (l_i, r_i]|\boldsymbol{\Theta}^s)\right)\right],$$

com $V_{s=1}^{S}\left[\log\left(P(T_i \in (l_i, r_i]|\boldsymbol{\Theta}^s)\right)\right]$ sendo a variância amostral, em que

$$V_{s=1}^{S} a^s = \frac{1}{S-1}\sum_{s=1}^{S}\left(a^s - \bar{a}\right),$$

de acordo com Gelman et al. (2013). Assim, temos que o critério $WAIC$ é estimado utilizando a seguinte expressão

$$\widehat{WAIC} = \widehat{\text{lpd}} - \widehat{\text{pd}}.$$

Maior valor de $WAIC$ indica que o modelo é preferível aos outros modelos.

Para facilitar a apresentação dos resultados dos modelos, apresentamos os critérios na mesma escala, multiplicando os valores de $LPML$ e $WAIC$ por -2 e dessa forma avaliamos o melhor modelo com o menor valor de critério (Gelman et al., 2013).

Capítulo 5

Aplicações

Com o intuito de ilustrar os modelos propostos no Capítulo 4, apresentamos nesse capítulo a análise de bancos de dados simulados e bancos de dados reais sem e com fração de cura, respectivamente. Sobre a análise de dados sem fração de cura, aplicamos os modelos a dados um conjunto de dados simulados e aos dados de câncer de mama previamente analisados por Finkelstein (1986) e Sinha et al. (1999). Na análise de dados com fração de cura, ilustramos os modelos com fração de cura dinâmicos através de um conjunto de dados simulados e para os dados reais utilizamos os dados de tempo até a soroconversão de HIV em pacientes hemofílicos (Goedert et al., 1989; Kroner et al., 1994). Para a obtenção dos intervalos de tempo para cada observação sujeita a censura intervalar nos dados simulados, utilizamos o mesmo algoritmo descrito em da Costa (2016) e os códigos para a geração dos dados está disponíveis no Apêndice B. * Conforme sugerido por Wang et al. (2013), duas formas de construção para a grade mais fina $\tau^{'}$ foram consideradas. Denotamos a grade do Tipo 1 a grade que é formada pelos limites inferiores e superiores distintos dos intervalos de tempo observados, enquanto a grade do Tipo 2 é determinada por intervalos equidistantes. Tanto nos dados reais quanto nos dados simulados os modelos foram comparados através dos critérios de seleção $LPML$, DIC e $WAIC$, descritos na Seção 4.3.

Para os modelos que utilizam a estrutura de agrupamento do MPP proposta por Demarqui et al. (2008), assumimos que $c_{I_\rho^{(j)}} = 1$, $j = 1, \ldots, b$, por não contarmos com nenhuma informação acerca da similaridade entre os intervalos. Dessa forma, a distribuição *a priori* para a partição aleatória ρ será do tipo Bayes-Laplace, ou seja, uma distribuição uniforme discreta.

Toda a modelagem proposta nesta tese foi implementada em linguagem `R` (R Core Team, 2016), versão 3.2.2, assim como a organização dos resultados. Usamos a função `ars` do pacote `ars` (Rodriguez, 2014) para gerar amostra de distribuições com função densidade de probabilidade log-côncavas e a função `arms` disponível no pacote `dlm` (Petris, 2010) para obter amostras de distribuições com funções densidade que não são log-côncavas. O pacote `dynsurv` (Wang et al., 2014) foi utilizado para ajustar os modelos propostos em Sinha et al. (1999) e Wang et al. (2013). Para a obtenção das curvas de sobrevivência através

do estimador de Turnbull (1976), foi utilizado o pacote `interval` (Fay & Shaw, 2010). Os testes de convergência das cadeias, Geweke (1992) e Heidelberger & Welch (1983), disponíveis no pacote `coda` (Plummer et al., 2006), foram utilizados para verificar a convergência das cadeias geradas, apresentando resultados satisfatórios em todos os casos.

5.1 Análise de dados sem fração de cura

Nesta seção apresentamos os resultados referentes aos modelos ajustados aos dados simulados e a dados reais sem fração de cura. Para facilitar a exposição dos resultados, utilizamos a seguinte nomenclatura para os modelos ajustados:

$\mathcal{M}_0$: MEP com grade fixa e coeficientes de regressão fixo no tempo (Sinha et al., 1999);

$\mathcal{M}_1$: MEP com grade fixa e coeficientes de regressão variando no tempo (Sinha et al., 1999);

$\mathcal{M}_2$: MEP com grade aleatória e coeficientes de regressão variando no tempo (Wang et al., 2013);

$\mathcal{M}_3$: MEP com grade aleatória e coeficientes de regressão fixo no tempo;

$\mathcal{M}_4$: MEP dinâmico com grade fixa e coeficientes de regressão variando no tempo;

$\mathcal{M}_5$: MEP dinâmico com grade aleatória e coeficientes de regressão variando no tempo,

em que os modelos $\mathcal{M}_3, \mathcal{M}_4$ e $\mathcal{M}_5$ representam os modelos propostos nesta tese.

5.1.1 Dados simulados

Nesta subseção apresentamos o algoritmo para a geração dos dados de sobrevivência sujeitos a censura intervalar em que a função risco tem a seguinte forma

$$h(t) = h_0(t) \exp\left(x\beta(t)\right).$$

Dessa forma, pelo método da transformação inversa temos que os valores dos tempos de falha são obtidos através da solução da seguinte expressão

$$u - H_0(t) \exp\left(x\beta(t)\right) = 0, \tag{5.1}$$

em que $u \sim Unif(0,1)$.

Assumimos uma distribuição Weibull para modelar a função risco de base, ou seja, $h_0(t) = \alpha\gamma t^{\alpha-1}$, em que α e γ são os parâmetros de forma e escala. Para garantir que o efeito da covariável varie no tempo utilizamos a seguinte função

$$\beta(t) = \begin{cases} 0,6 & , \text{ se } \quad t \leq 1,5, \\ 1,4 & , \text{ se } \quad t > 1,5, \end{cases}$$

e $x \sim$ Bernoulli(0.5). De acordo com da Costa (2016), o algoritmo para a geração dos dados de sobrevivência com censura intervalar é dado por:

1. Obtenha uma amostra de tamanho n de tempos de falha resolvendo numericamente a expressão (5.1).

2. Para a geração dos tempos censura assumimos que $C_i \sim$ Weibull(a, b) e obtenha $\delta_i = I(T_i \leq C_i)$, para $i = 1, \ldots, n$.

3. Para construir os intervalos de tempo, assuma que $L_i = C_i$ e $R_i = \infty$, se $\delta_i = 0$, representando um caso de censura a direita.

4. Caso $\delta_i = 1$, gere V tempos de monitoramento a partir da soma de unidades com distribuição uniforme $U_g \sim Unif(d_1, d_2)$, para $g = 1, \ldots, V$. Assuma que o primeiro tempo de monitoramento seja igual a 0, ou seja, $U_0 = 0$. Assim os intervalos de tempo são obtidos tomando $L_i = \sum_{g=0}^{V-1} U_g$ e $R_i = \sum_{g=0}^{V} U_g$.

Para esta análise estipulamos um tamanho amostral igual a 500 observações. Os valores reais dos parâmetros utilizados na geração dos dados são os seguintes:

- Para a distribuição dos tempos de falha os valores dos parâmetros da distribuição Weibull são: $\alpha = 1,5$ e $\gamma = 0,3$;

- E para a distribuição dos tempos de censura os valores dos parâmetros da distribuição Weibull foram iguais a, $a = 1,5$ e $b = 0,1$.

- Os valores de $d_1 = 0,1$ e $d_2 = 0,5$.

Os valores especificados para as distribuições *a priori* foram escolhidos para representar a informação vaga acerca dos parâmetros de interesse. Assim, as especificações *a priori* para os modelos nesta aplicação foram:

$\mathcal{M}_0$ e $\mathcal{M}_1$: Os valores dos hiperparâmetros da distribuição gama são: $\alpha_j = 0.01$ e $\gamma_j = 0.01$, para $j = 1, \ldots, b$. Para o modelo $\mathcal{M}_0$, assumimos uma distribuição normal para cada componente do vetor de coeficientes de regressão, com média 0 e variância $\omega^2 = 1000$. Para o modelo $\mathcal{M}_1$, assumimos média 0 somente para $j = 1$ uma vez que para $j \geq 2$ é utilizado um passeio aleatório, e variância $\omega_j^2 = 1000$, para $j = 1, \ldots, b$.

$\mathcal{M}_2$: Neste modelo assumimos uma distribuição normal *a priori* para cada componente do vetor $\theta(t)$, uma distribuição *a priori* gama invertida, com hiperparâmetros de forma e escala iguais a 2 e 1, respectivamente, assim como em (4.4) , o incremento na variância do primeiro intervalo $a_0 = 1$ e uma distribuição uniforme discreta para o número de saltos b, ou seja, $p(b) \propto 1/m'$, em que m' é o número de intervalos associados a partição inicial.

$\mathcal{M}_3$**:** As especificações para os coeficientes de regressão e para o vetor das taxas do MEP foram mesma que no modelo $\mathcal{M}_0$.

$\mathcal{M}_4$ **e** $\mathcal{M}_5$**:** Temos as seguintes especificações iniciais,

$$\mathbf{m}_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{e} \quad \mathbf{C}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$\mathbf{G}_{\rho}^{(j)} = \mathbf{I}$, para $j = 1, \ldots, b$. Para avaliar a sensibilidade de ϕ nos ajustes consideramos uma malha de 7 valores, 0.1, 0.2, 0.4, 0.5, 0.6, 0.8, 0.9.

Comparamos os modelos $\mathcal{M}_0$ a $\mathcal{M}_5$ utilizando somente para a grade inicial do Tipo 1 e valores de $m^{'}$ iguais a 10, 20 e 30. A análise bayesiana contou com um total de 100000 iterações, com um preaquecimento da cadeia de 50000 iterações, um espaçamento de 10, resultando em uma amostra *a posteriori* de tamanho 5000.

A Tabela 5.1 fornece os valores dos critérios de seleção para os modelos $\mathcal{M}_0$ a $\mathcal{M}_3$. É possível observar que, o modelo $\mathcal{M}_2$ se sobressai quando comparado aos demais modelos ajustados, com valores de LPML igual a 1943,27, $WAIC$ da ordem de 1949,25 e DIC igual a 1944,01, para $m^{'} = 30$. Um ponto importante a ressaltar é que para alguns valores de $m^{'}$, principalmente para o $WAIC$, os modelos apresentaram alguns problemas para obter as funções de verossimilhança avaliadas nas amostras *a posteriori*, impossibilitado o cálculo dos critérios. Vale ressaltar que os modelos com problemas, foram ajustados com a ajuda do pacote `dynsurv`.

Tabela 5.1: Critérios $LPML$, $WAIC$ e DIC para os modelos $\mathcal{M}_0$ a $\mathcal{M}_3$.

Critérios	Modelos	$m^{'}$		
		10	20	30
$LPML$	M_0	2634,10	-	2598,22
	M_1	2510,40	-	2498,97
	M_2	1959,27	-	1943,88
	M_3	2533,14	2503,00	2505,09
$WAIC$	M_0	2700,04	-	-
	M_1	2582,40	-	-
	M_2	1961,02	-	1949,25
	M_3	2661,31	2619,41	2623,69
DIC	M_0	4422,55	-	4228,57
	M_1	3828,87	-	3247,11
	M_2	1970,74	-	1944,01
	M_3	4069,20	3926,20	3791,01

Para os modelos $\mathcal{M}_4$ e $\mathcal{M}_5$, observamos nas Tabelas 5.2, 5.3 e 5.4 os critérios de seleção $LPML$, $WAIC$ e DIC, respectivamente. Observamos que o modelo $\mathcal{M}_4$ apresentou melhores resultados quando comparado ao modelo $\mathcal{M}_5$, com os valores de $LPML$, $WAIC$ e DIC iguais a $1953,92$, $1953,62$ e $1963,39$, respectivamente. Com esses resultados o modelo indicado é $\mathcal{M}_4$ com $m^{'} = 30$ e fator de desconto igual a 0,2. É importante observar que tanto o modelo $\mathcal{M}_2$ quanto o modelo $\mathcal{M}_4$ apresentam os melhores valores de critérios, que $\mathcal{M}_2$ aparece com um ajuste aos dados um pouco melhor.

Tabela 5.2: Critério $LPML$ para os modelos $\mathcal{M}_4$ e $\mathcal{M}_5$.

Fator	LPML					
de	τ fixo	τ aleatório	τ fixo	τ aleatório	τ fixo	τ aleatório
desconto	$m^{'} = 10$		$m^{'} = 20$		$m^{'} = 30$	
0,1	1964,76	2021,01	1955,90	1987,31	1955,22	1974,66
0,2	1969,93	2027,77	1960,51	1990,60	1953,92	1976,38
0,4	1975,68	2036,76	1964,04	1995,18	1955,13	1977,52
0,5	1984,33	2051,05	1967,53	2003,74	1957,69	1982,26
0,6	1997,78	2066,90	1972,06	2014,45	1960,90	1989,68
0,8	2018,11	2086,73	1979,71	2031,34	1965,60	2001,92
0,9	2134,57	2164,21	2087,33	2135,45	2053,71	2010,88

Tabela 5.3: Critério $WAIC$ para os modelos $\mathcal{M}_4$ e $\mathcal{M}_5$.

Fator	WAIC					
de	τ fixo	τ aleatório	τ fixo	τ aleatório	τ fixo	τ aleatório
desconto	$m^{'} = 10$		$m^{'} = 20$		$m^{'} = 30$	
0,1	1964,68	2021,00	1955,57	1987,12	1954,53	1974,41
0,2	1969,90	2027,81	1960,44	1990,50	1953,62	1976,22
0,4	1975,67	2036,78	1964,01	1995,10	1955,02	1977,42
0,5	1984,33	2051,08	1967,53	2003,73	1957,65	1982,22
0,6	1997,78	2066,92	1972,06	2014,45	1960,88	1989,66
0,8	2018,11	2086,75	1979,71	2031,34	1965,60	2001,91
0,9	2134,57	2164,21	2087,33	2135,45	2053,71	2010,01

Tabela 5.4: Critério DIC para os modelos $\mathcal{M}_4$ e $\mathcal{M}_5$.

Fator	DIC					
de	τ fixo	τ aleatório	τ fixo	τ aleatório	τ fixo	τ aleatório
desconto	$m^{'} = 10$		$m^{'} = 20$		$m^{'} = 30$	
0,1	1974,92	2755,21	1966,74	2127,68	1964,19	2028,10
0,2	1978,59	2749,89	1971,25	2148,50	1963,39	2040,20
0,4	1983,27	2760,83	1972,86	2186,29	1964,42	2042,65
0,5	1990,86	2758,23	1975,20	2238,46	1965,93	2066,19
0,6	2003,98	2702,57	1978,82	2283,00	1967,68	2109,44
0,8	2022,94	2624,76	1985,23	2345,74	1971,69	2163,98
0,9	2136,65	2234,30	2089,57	2236,42	2056,24	2200,14

Apresentamos nas Figuras 5.1 a 5.6 uma comparação das funções de sobrevivências estimadas pelos modelo ajustados, pelo método de Turnbull e da função de sobrevivência real. Assim como observamos na análise dos critérios, os modelo $\mathcal{M}_2$, $\mathcal{M}_4$ e $\mathcal{M}_5$ acompanham bem o comportamento da curvas de sobrevivência via Turnbull e real.

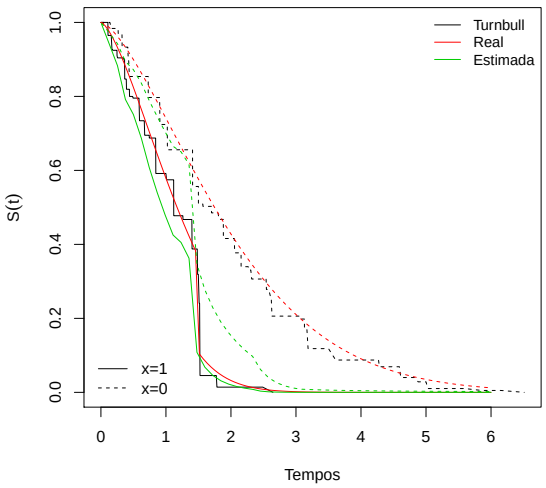

Figura 5.1: Estimativa da função de sobrevivência para o modelo $\mathcal{M}_0$

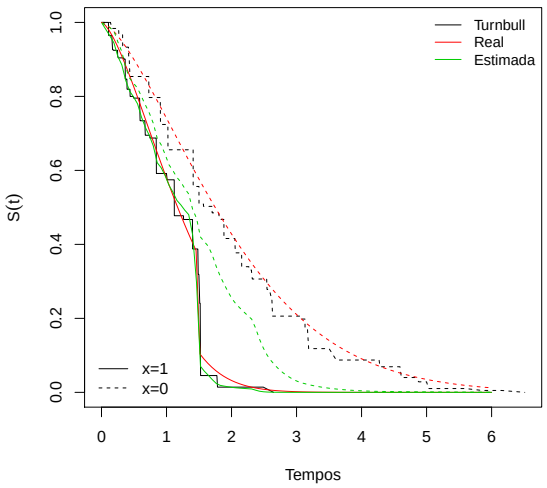

Figura 5.2: Estimativa da função de sobrevivência para o modelo $\mathcal{M}_1$

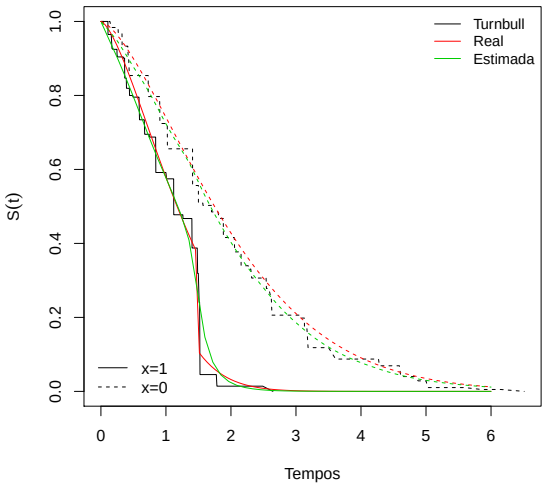

Figura 5.3: Estimativa da função de sobrevivência para o modelo $\mathcal{M}_2$

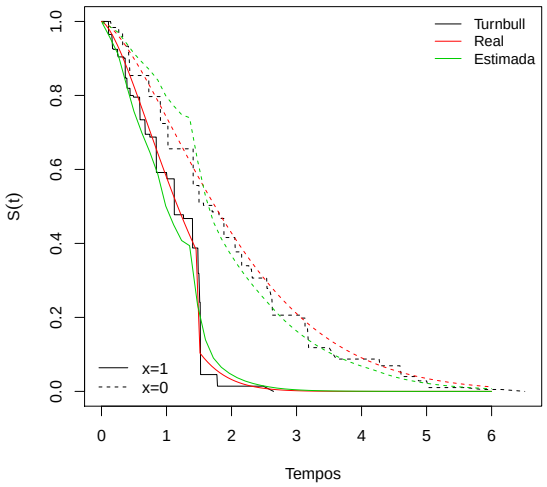

Figura 5.4: Estimativa da função de sobrevivência para o modelo $\mathcal{M}_3$

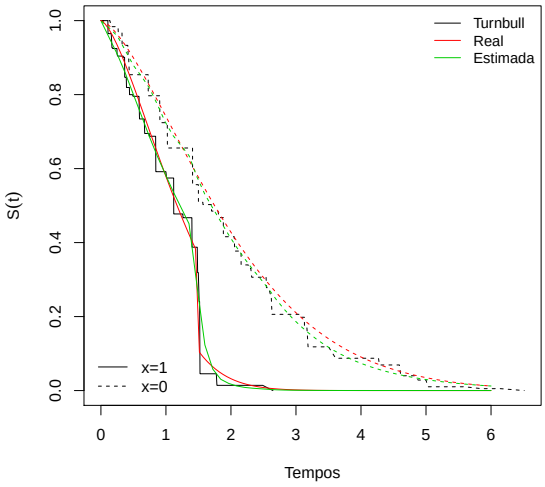

Figura 5.5: Estimativa da função de sobrevivência para o modelo $\mathcal{M}_4$

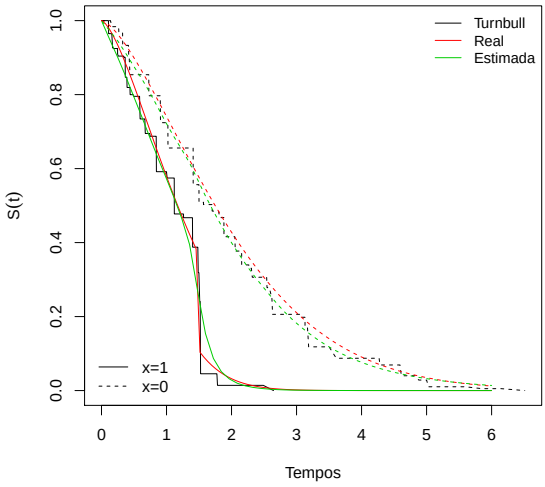

Figura 5.6: Estimativa da função de sobrevivência para o modelo $\mathcal{M}_5$

As Figuras 5.7 a 5.10 representam a média *a posteriori* de β_1 para o modelo comparando o efeito real do coeficiente. Observamos que, todos os modelos com coeficientes de regressão que variam no tempo, conseguem acompanhar o comportamento real do efeito ao longo do tempo, antes de após o ponto de mudança.

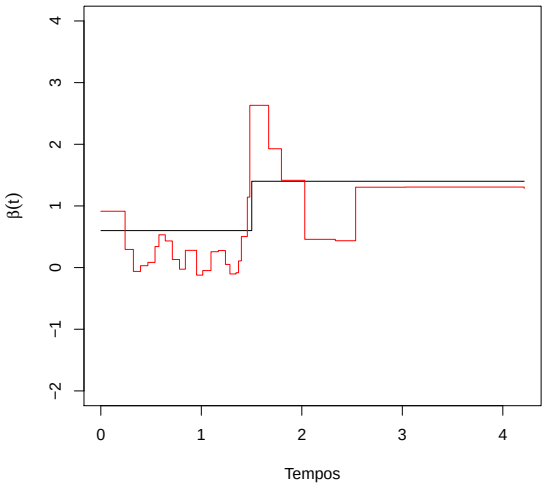

Figura 5.7: Média *a posteriori* de β_1 comparado ao efeito real para o modelo $\mathcal{M}_1$.

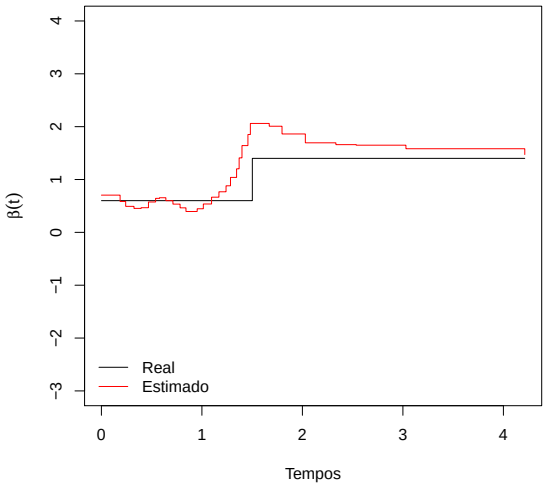

Figura 5.8: Média *a posteriori* de β_1 comparado ao efeito real para o modelo $\mathcal{M}_2$.

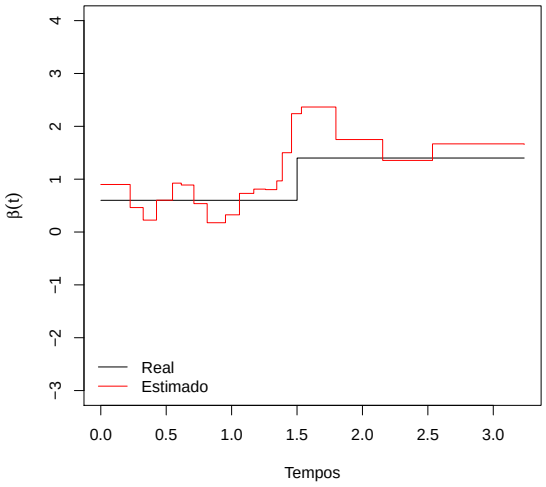

Figura 5.9: Média *a posteriori* de β_1 comparado ao efeito real para o modelo $\mathcal{M}_4$.

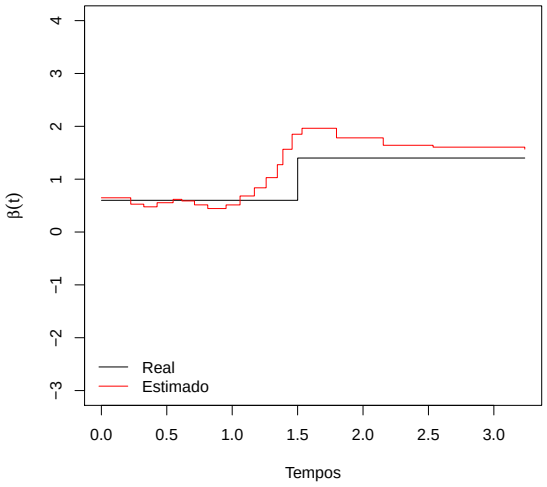

Figura 5.10: Média *a posteriori* de β_1 comparado ao efeito real para o modelo $\mathcal{M}_4$.

Na Tabela 5.5 apresentamos as estatísticas descritivas sobre a amostra *a posteriori* do número de intervalos estimados pelos modelos com grade aleatória. Podemos notar que o modelo $\mathcal{M}_3$ apresentou o maior número de intervalos *a posteriori* com uma moda igual a 24 e com intervalo HPD 95% de $[20; 27]$.

Tabela 5.5: Estatísticas descritivas para a amostra *a posteriori* do número de intervalos.

Modelos	Mínimo	Máximo	Média	Mediana	Moda	Desvio-padrão	HPD 95%
$\mathcal{M}_2$	6	24	15,57	16	15	2,66	[10;20]
$\mathcal{M}_3$	16	30	24,25	24	24	1,93	[20;27]
$\mathcal{M}_{5_{\phi=0,2}}$	4	18	10,56	11	11	2,14	[6;14]

5.1.2 Dados de câncer de mama

Nesta subseção apresentamos uma análise a dados reais, em que utilizamos os dados de câncer de mama inicialmente analisados em Finkelstein (1986), que têm sido extensivamente utilizados para ilustrar inúmeras propostas de modelagem para dados de sobrevivência com censura intervalar (veja, Sinha et al. (1999); Goetghebeur & Ryan (2000); Pan (2000); Chen et al. (2013)), entre outros. O objetivo do estudo era avaliar o estado da paciente e um dos quesitos incluía a deterioração da mama. O tempo até a deterioração da mama das pacientes é a variável resposta de interesse. Devido às consultas com o especialista ocorrerem entre 4 a 6 semanas e não sabermos quando houve exatamente a deterioração, mas somente o intervalo de tempo da ocorrência, os dados são caracterizados pela presença de censura intervalar.

O dados contam com um total de 94 pacientes, dos quais aproximadamente 60% tiveram a mama deteriorada. As pacientes foram divididas em dois grupos para comparação: 46 pacientes que receberam somente radioterapia ($x = 0$) e 48 pacientes que receberam radioterapia mais quimioterapia ($x = 1$).

Assim como em Sinha et al. (1999), para os modelos $\mathcal{M}_0$ e $\mathcal{M}_1$ especificamos os valores de hiperparâmetros $\alpha_j = 0.2$ e $\gamma_j = 0.4$, para $j = 1, \ldots, b$. Para os coeficientes de regressão, no modelo $\mathcal{M}_0$, assumimos uma distribuição normal para cada componente do vetor de coeficientes, com média 0 e variância $\omega^2 = 1000$. Similarmente, para o modelo $\mathcal{M}_1$, assumimos média 0 para o primeiro intervalo e um passeio aleatório para os demais com variância $\omega_j^2 = 1000$, para $j = 1, \ldots, b$.

Para o modelo $\mathcal{M}_2$, assumimos uma distribuição normal *a priori* para cada componente do vetor $\theta(t)$ induzido pelos pontos de saltos, uma distribuição *a priori* gama invertida para ω^2, com hiperparâmetros de escala e locação iguais a 2 e 1, respectivamente, e $a_0 = 1$ sendo o valor de controle da variabilidade no primeiro intervalo no mesmo formato como apresentamos na expressão (4.4) na Subseção 4.1.2. Para os pontos de saltos tomamos uma distribuição uniforme discreta para o número de saltos b, ou seja, $p(b) \propto 1/m'$. Para o modelo $\mathcal{M}_3$, foram utilizadas as mesmas especificações *a priori* do modelo $\mathcal{M}_0$.

Finalmente, para os modelos $\mathcal{M}_4$ e $\mathcal{M}_5$, as especificações para os momentos dos coeficientes de regressão foram as seguintes:

$$\mathbf{m}_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{e} \quad \mathbf{C}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Com o objetivo de avaliar o impacto da especificação do fator de desconto ϕ sobre a qualidade do ajuste dos modelos, foi realizada uma análise de sensibilidade considerando uma malha de dezoito valores para ϕ, variando de $0,10$ a $0,95$, com espaçamento de 0.05. Também foi assumido que $\mathbf{G}_\rho^{(j)} = \mathbf{I}$, para $j = 1, \ldots, b$. Para todos os modelos, avaliamos os valores de m' iguais a 10, 20, 30 e 41, para os dois tipos especificações de grade descritos anteriormente. Para cada modelo ajustado foi considerado apenas uma cadeia de 100000 iterações, com 50000 descartadas no aquecimento da cadeia, com um espaçamento de 10,

resultando em uma amostra *a posteriori* de tamanho 5000.

A Tabela 5.6 mostra os valores de $LPML$, DIC e $WAIC$ para os modelos $\mathcal{M}_0$ e $\mathcal{M}_1$. Ao comparar $\mathcal{M}_0$ e $\mathcal{M}_1$, observamos que o modelo que melhor se ajustou aos dados de câncer de mama foi $\mathcal{M}_1$, apresentando um valor de $LPML$ igual a 299,73, $WAIC$ igual a 299,01 e DIC igual a 292,22, para $m^{'} = 41$ e grade do Tipo 2.

Similarmente, para o modelo $\mathcal{M}_0$, o melhor tipo de grade foi o Tipo 2 e $m^{'} = 41$. Veja que dois dos três critérios de comparação de modelos, $WAIC$ e DIC, com valores iguais a 302,76 e 297,80, apontaram esta configuração, apesar de somente o critério $LPML$, com valor igual a 304,71, indicar a grade do Tipo 1, para ambos os modelos, os critérios estão indicando que um maior número de intervalos iniciais leva a um melhor qualidade de ajuste do modelo aos dados. Similar aos ajustes com dados simulados, alguns valores dos critérios de seleção não puderam ser obtidos.

Tabela 5.6: Critérios $LPML$, $WAIC$ e DIC para os modelos $\mathcal{M}_0$ e $\mathcal{M}_1$.

Tipos de $m^{'}$	$LPML$				$WAIC$				DIC			
	$\mathcal{M}_0$											
	$m=10$	$m=20$	$m=30$	$m=41$	$m=10$	$m=20$	$m=30$	$m=41$	$m=10$	$m=20$	$m=30$	$m=41$
Tipo 1	-	352,08	305,81	304,71	-	-	-	302,82	-	308,78	300,30	298,91
Tipo 2	-	317,34	310,02	304,74	-	-	-	302,76	-	308,77	301,63	297,80
	$\mathcal{M}_1$											
Tipo 1	-	348,97	303,16	302,91	-	-	-	301,32	-	303,73	295,09	294,03
Tipo 2	-	315,99	306,15	299,73	-	-	-	299,01	-	302,85	295,71	292,22

Na Tabela 5.7, os resultados de $LPML$, $WAIC$ e DIC indicam para o modelo $\mathcal{M}_2$ que a melhor configuração para modelar os dados de câncer é através de uma grade do Tipo 2 com $m^{'} = 30$, devido aos valores de $LPML$ e DIC de 289,72 e 288,92, respectivamente. Mais uma vez, segundo os critérios $LPML$ e DIC, há evidências em não utilizar valores pequenos para $m^{'}$. Assim como na análise de dados simulados, houveram problemas no cálculo da função de verossimilhança avaliadas nas amostras *a posteriori* e consequentemente no cálculo dos critérios.

Tabela 5.7: Critérios $LPML$, $WAIC$ e DIC para o modelo $\mathcal{M}_2$.

Tipos de $m^{'}$	$LPML$				$WAIC$				DIC			
	$m=10$	$m=20$	$m=30$	$m=41$	$m=10$	$m=20$	$m=30$	$m=41$	$m=10$	$m=20$	$m=30$	$m=41$
Tipo 1	-	295,50	294,32	291,81	-	-	-	292,00	-	294,76	293,25	290,79
Tipo 2	-	294,80	289,72	291,47	-	-	-	291,54	-	293,84	288,92	290,88

Podemos observar na Tabela 5.8 os resultados dos critérios de seleção para o modelo $\mathcal{M}_3$. Observamos que os valores de $LPML$, $WAIC$ e DIC iguais a 304,96, 304,43 e 314,74, respectivamente, nos indicam que a melhor configuração para o modelo é obtida quando $m^{'} = 10$ e grade do Tipo 1.

Tabela 5.8: Critérios $LPML$, DIC e $WAIC$ para o modelo $\mathcal{M}_3$.

Tipos de $m^{'}$	$LPML$				$WAIC$				DIC			
	$m=10$	$m=20$	$m=30$	$m=41$	$m=10$	$m=20$	$m=30$	$m=41$	$m=10$	$m=20$	$m=30$	$m=41$
Tipo 1	304,96	317,15	314,92	315,60	304,43	314,55	311,79	311,79	314,74	399,89	361,59	348,59
Tipo 2	317,00	324,34	322,66	317,64	314,09	319,92	318,73	314,89	508,33	555,90	457,02	369,81

Devido ao problema com o cálculo dos critérios, principalmente para os valores de $WAIC$, não comparamos os modelos $\mathcal{M}_0$, $\mathcal{M}_1$, $\mathcal{M}_2$ e $\mathcal{M}_3$ através de figuras para o mesmo, somente no formato de tabelas.

Nas Figuras 5.11 e 5.12 apresentamos a comparação entre os modelos $\mathcal{M}_0$ a $\mathcal{M}_3$ discutidos. Observe que, em ambos os gráficos há a presença de um comportamento monótono para os dois tipos de grade, em que a medida que $m^{'}$ aumenta, a qualidade de ajuste melhora, exceto para o modelo $\mathcal{M}_3$. Note também que o modelo $\mathcal{M}_2$ apresenta um melhor desempenho quando comparado com os modelos $\mathcal{M}_0$, $\mathcal{M}_1$ e $\mathcal{M}_3$.

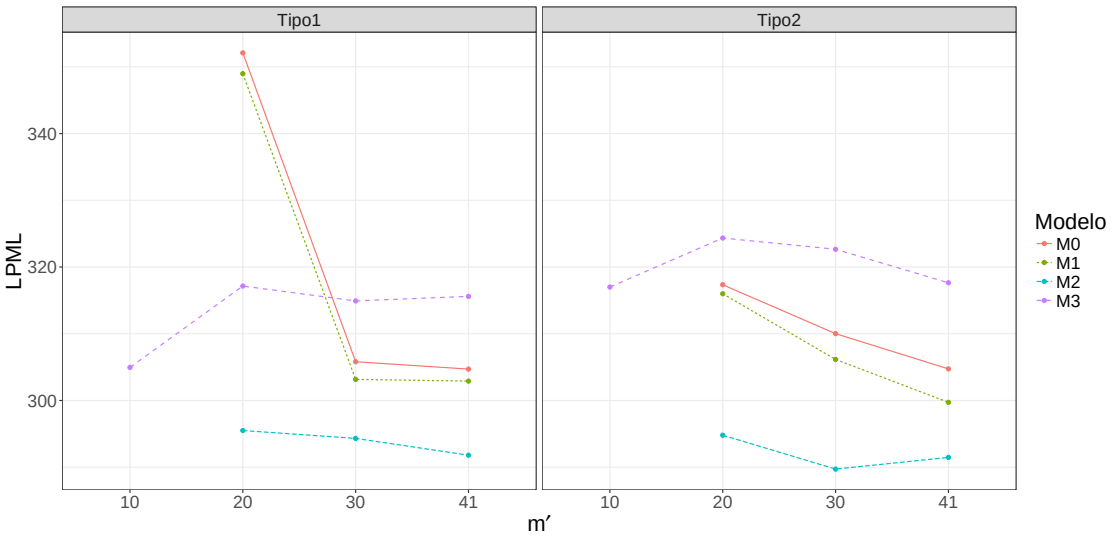

Figura 5.11: Comparação dos modelos $\mathcal{M}_0$,$\mathcal{M}_1$, $\mathcal{M}_2$ e $\mathcal{M}_3$ via $LPML$, para os diferentes valores de $m^{'}$ e tipos de grade.

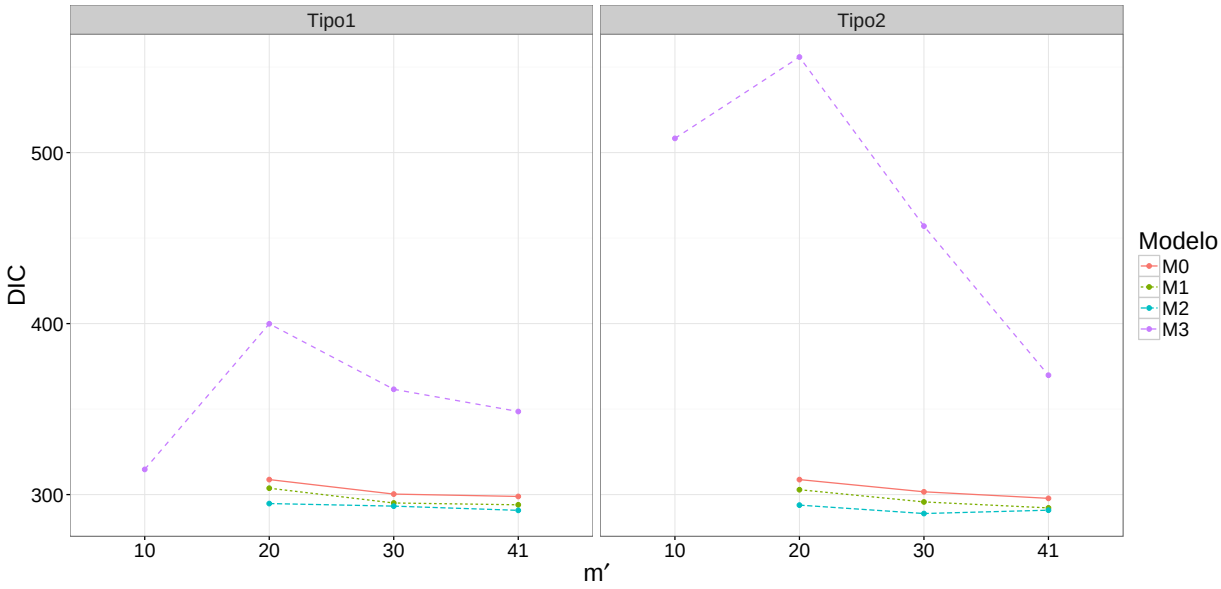

Figura 5.12: Comparação dos modelos $\mathcal{M}_0$, $\mathcal{M}_1$, $\mathcal{M}_2$ e $\mathcal{M}_3$, via DIC, para os diferentes valores de $m^{'}$ e tipos de grade.

Finalmente apresentamos os resultados dos critérios de seleção de modelos $LPML$, $WAIC$ e DIC para os modelos $\mathcal{M}_4$ e $\mathcal{M}_5$, nas Tabelas 5.9, 5.10 e 5.11, respectivamente. Observamos que o melhor modelo $\mathcal{M}_4$ é com grade do Tipo 1, $m^{'} = 20$ e com $\phi = 0,7$, com valores de $LPML$, $WAIC$ iguais a 286,92 e 268,85, respectivamente. Para o modelo $\mathcal{M}_5$, os valores de $LPML$ e $WAIC$ são iguais a 291.25 e 291.14, respectivamente, para a grade do tipo 1 e $m^{'} = 20$ para um valor de ϕ igual a 0.6. Dessa forma, o modelo $\mathcal{M}_{4_{\phi=0.7}}$ apresenta melhor qualidade de ajuste aos dados quando comparado ao modelo $\mathcal{M}_{5_{\phi=0.6}}$.

Tabela 5.9: Valores dos critérios $LPML$ para os modelos $\mathcal{M}_4$ e $\mathcal{M}_5$ para grade do Tipo 1.

ϕ	$LPML$							
	$m^{'} = 10$		$m^{'} = 20$		$m^{'} = 30$		$m^{'} = 41$	
	τ fixo	τ aleatório	τ fixo	τ aleatório	τ fixo	τ aleatório	τ fixo	τ aleatório
				Grade tipo 1				
0,1	301,75	312,71	309,48	304,23	315,04	309,93	314,00	310,75
0,15	298,53	300,02	307,12	301,45	314,17	304,76	317,87	308,90
0,2	296,17	294,02	303,97	299,58	310,95	303,20	318,13	306,39
0,25	296,79	292,54	301,24	297,16	306,68	299,78	311,25	303,37
0,3	292,20	296,19	297,37	296,05	306,59	298,57	307,64	300,49
0,35	291,21	293,99	295,49	294,43	302,36	296,50	306,06	298,69
0,4	290,29	293,74	293,23	293,18	299,73	295,30	304,51	296,72
0,45	289,72	293,70	291,63	292,50	297,98	294,44	300,87	296,12
0,5	289,24	293,67	290,10	291,96	295,93	293,29	299,13	294,43
0,55	289,00	294,06	289,02	291,39	294,94	293,27	297,12	293,63
0,6	288,96	294,47	287,94	291,25	293,76	292,67	294,85	291,99
0,65	289,48	294,97	287,33	291,30	293,13	292,61	293,24	292,05
0,7	290,18	295,77	286,92	291,66	292,66	292,45	292,24	291,60
0,75	291,16	296,58	286,99	292,36	292,42	293,48	291,40	291,55
0,8	292,56	297,47	287,68	293,46	292,07	294,27	291,46	292,45
0,85	294,38	298,68	289,04	294,87	293,65	295,87	291,11	293,80
0,9	296,53	299,71	291,57	296,69	297,01	297,89	293,48	296,03
0,95	299,08	301,17	295,46	299,12	300,95	301,12	297,45	300,67

Tabela 5.10: Valores dos critérios $WAIC$ para os modelos $\mathcal{M}_4$ e $\mathcal{M}_5$ para grade do Tipo 1.

ϕ	$WAIC$							
	$m^{'}=10$		$m^{'}=20$		$m^{'}=30$		$m^{'}=41$	
	τ fixo	τ aleatório	τ fixo	τ aleatório	τ fixo	τ aleatório	τ fixo	τ aleatório
	Grade tipo 1							
0,1	299,81	305,88	305,13	301,81	304,08	303,98	306,11	305,14
0,15	297,21	298,19	302,74	299,35	304,61	301,96	307,09	303,56
0,2	295,14	293,26	300,35	297,73	304,73	299,93	307,16	301,73
0,25	293,57	291,47	298,48	295,87	303,59	298,29	306,49	299,86
0,3	292,01	295,45	295,86	294,96	302,19	296,92	305,03	298,52
0,35	291,05	293,80	294,24	293,71	300,83	295,58	303,68	297,04
0,4	290,12	293,53	292,51	292,78	299,05	294,65	302,18	295,61
0,45	289,61	293,64	291,00	291,98	297,47	293,66	300,00	294,62
0,5	289,15	293,61	289,70	291,66	295,74	292,88	298,29	293,56
0,55	288,96	294,00	288,77	291,24	294,76	292,54	296,38	292,51
0,6	288,94	294,45	287,76	291,14	293,69	292,18	294,53	291,64
0,65	289,47	294,95	287,20	291,25	293,10	292,17	293,08	291,51
0,7	290,17	295,74	286,85	291,61	292,59	292,24	291,90	291,21
0,75	291,15	296,57	286,95	292,32	292,33	293,03	291,33	291,18
0,8	292,56	297,46	287,66	293,44	292,13	293,85	291,33	291,91
0,85	294,38	298,68	289,02	294,86	293,51	295,25	291,07	293,42
0,9	296,53	299,70	291,57	296,68	295,99	297,60	293,11	295,49
0,95	299,08	301,16	295,46	299,12	299,83	300,78	296,83	299,56

Tabela 5.11: Valores dos critérios DIC para os modelos $\mathcal{M}_4$ e $\mathcal{M}_5$ para grade do Tipo 1.

ϕ	DIC							
	$m^{'}=10$		$m^{'}=20$		$m^{'}=30$		$m^{'}=41$	
	τ fixo	τ aleatório	τ fixo	τ aleatório	τ fixo	τ aleatório	τ fixo	τ aleatório
	Grade tipo 1							
0,1	306,74	314,25	313,53	308,90	314,12	311,27	315,90	313,65
0,15	303,52	304,14	311,45	306,35	314,19	308,63	316,33	310,83
0,2	302,28	298,11	308,09	304,32	313,90	306,03	316,29	308,28
0,25	299,71	304,17	305,42	301,77	313,11	304,16	315,37	306,48
0,3	297,21	318,95	302,10	300,39	310,58	301,96	312,86	304,21
0,35	296,31	304,40	300,13	299,42	308,29	301,53	311,65	302,62
0,4	294,57	303,97	298,16	298,60	306,91	300,01	310,02	301,29
0,45	293,68	303,89	295,98	297,40	304,26	299,22	306,93	299,96
0,5	292,95	304,09	294,58	297,15	302,63	297,83	305,50	298,96
0,55	292,69	304,10	293,10	296,65	302,03	299,70	302,95	298,00
0,6	292,30	303,76	291,58	296,91	300,63	300,98	301,11	296,31
0,65	292,71	303,57	290,52	296,92	301,22	302,39	298,93	298,97
0,7	293,01	303,55	289,81	297,58	302,70	300,51	298,32	297,97
0,75	293,55	303,41	289,56	297,99	306,72	309,80	299,33	299,44
0,8	294,81	302,92	289,93	298,88	305,95	312,29	303,63	303,23
0,85	296,38	302,89	291,05	299,34	319,76	317,97	304,00	310,45
0,9	298,31	302,57	293,08	299,91	331,75	317,75	322,71	315,98
0,95	300,41	303,16	296,53	301,10	351,90	323,82	334,89	335,77

As Figuras 5.13 a 5.15 apresentam os valores dos critérios de seleção, $LPML$, $WAIC$ e DIC, para diferentes valores de fator de desconto, $m^{'}$ e tipos de grade para os modelos $\mathcal{M}_4$ e $\mathcal{M}_5$.

Observamos nas Figuras 5.13 a 5.15 que para todos os valores de $m^{'}$ e grade do Tipo 1, no modelo $\mathcal{M}_4$, a curva dos valores dos critérios apresentam um comportamento convexo, enquanto para o modelo $\mathcal{M}_5$, o comportamento é convexo somente para alguns valores de ϕ menores que 0.3, talvez por alguma instabilidade numérica em passar pouca informação entre os intervalos. Observe para a grade do Tipo 1 e $m^{'} = 10$, os critérios apresentaram instabilidade numérica para valores pequenos de ϕ, que nos dá um indicativo que é necessário passar mais informações entre os intervalos mesmo com $m^{'} = 10$, aumentando o valor de ϕ.

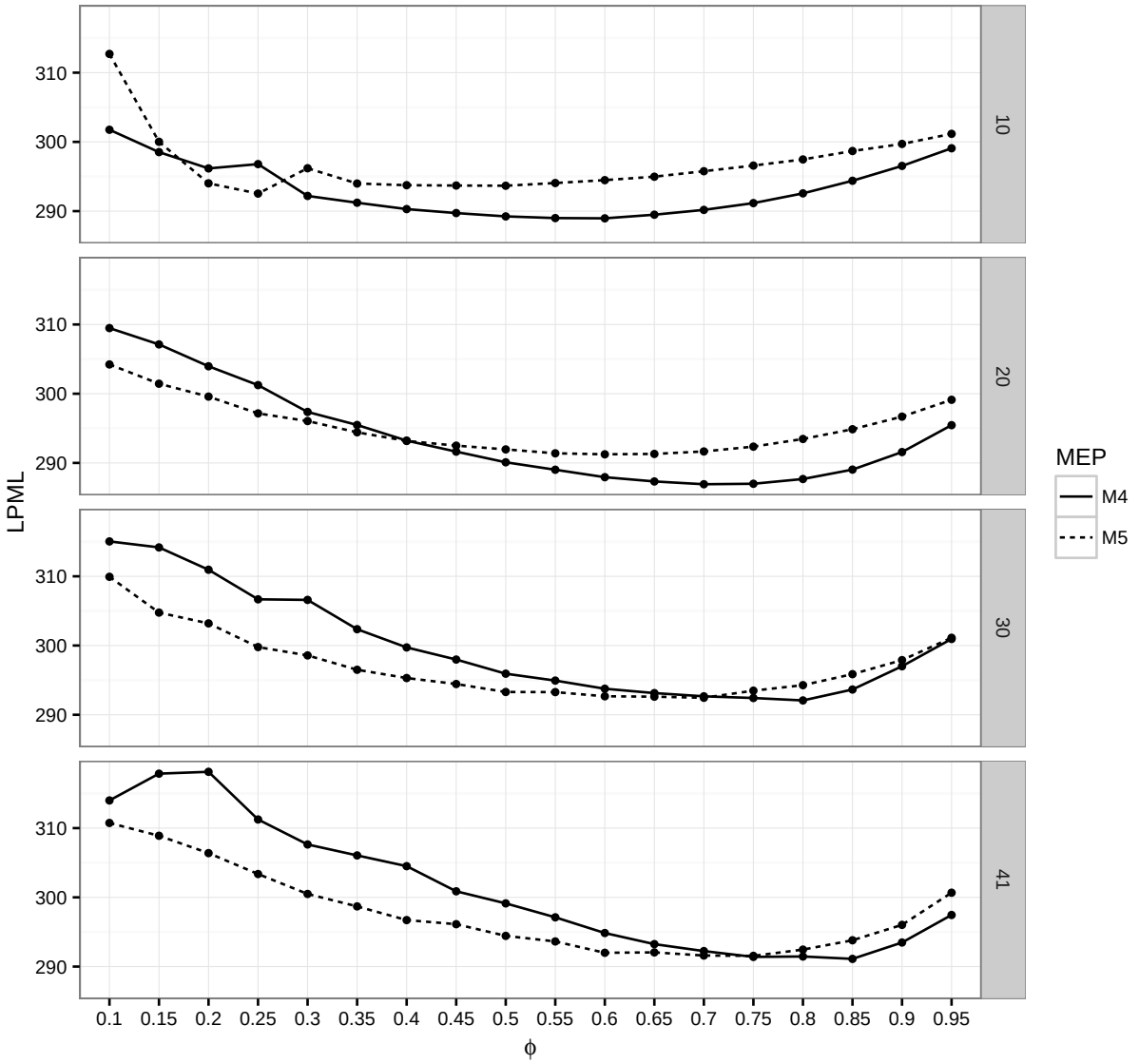

Figura 5.13: Comparação dos modelos $\mathcal{M}_4$ e $\mathcal{M}_5$, via $LPML$, para os diferentes valores ϕ, $m^{'}s$ e grade Tipo 1.

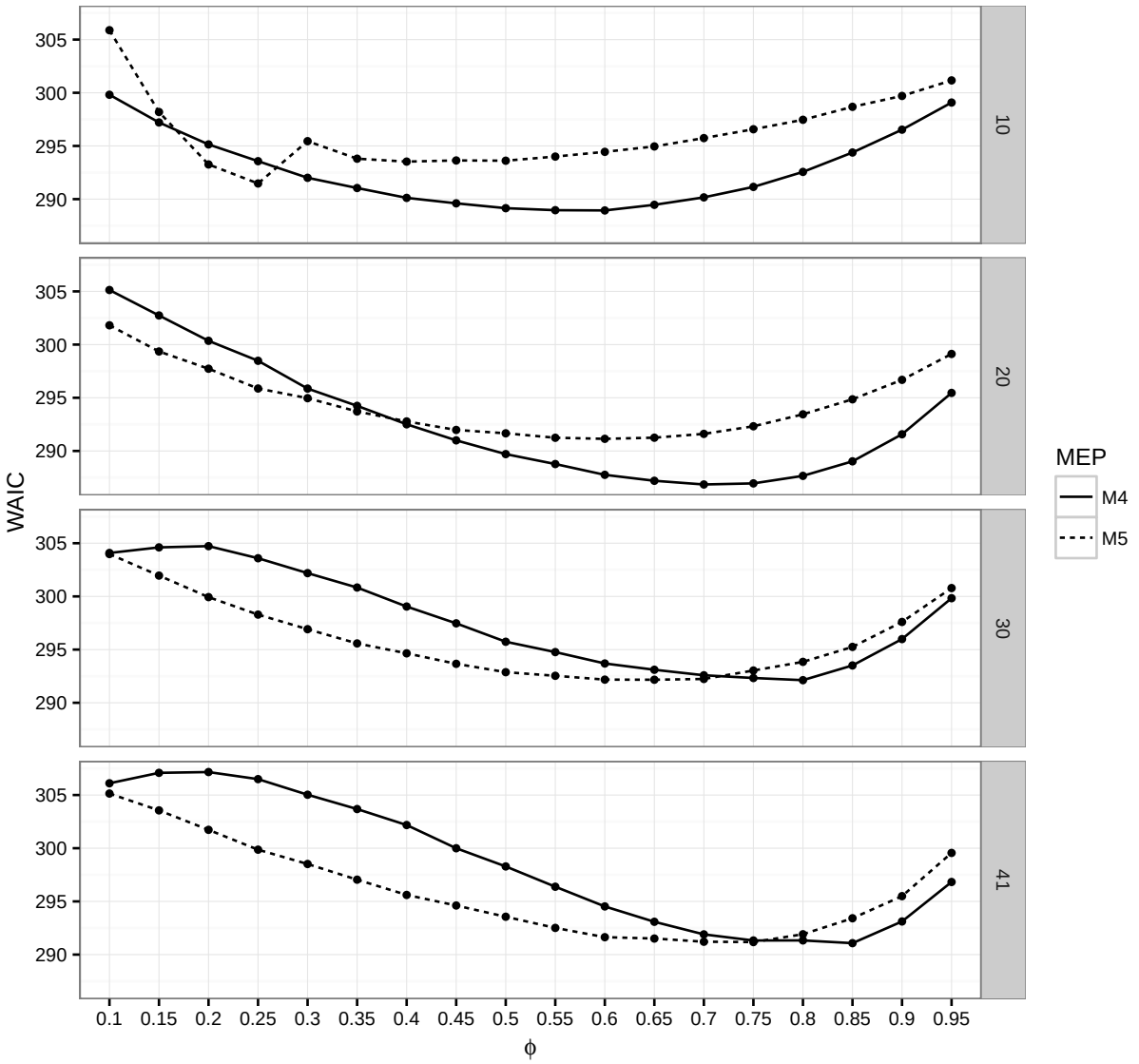

Figura 5.14: Comparação dos modelos $\mathcal{M}_4$ e $\mathcal{M}_5$, via $WAIC$, para os diferentes valores ϕ, $m^{'}s$ e grade Tipo 1.

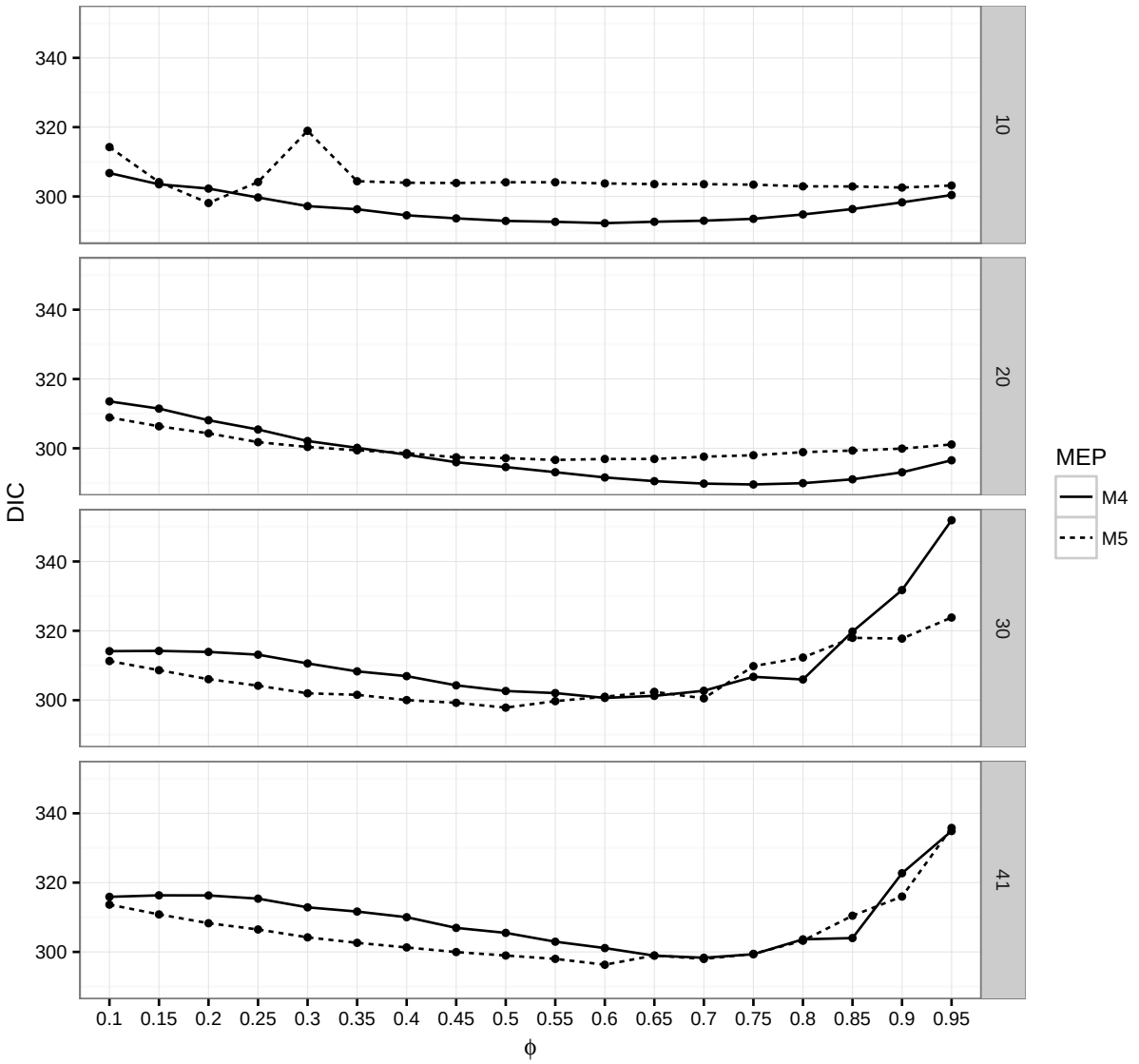

Figura 5.15: Comparação dos modelos $\mathcal{M}_4$ e $\mathcal{M}_5$, via DIC, para os diferentes valores ϕ, $m^{'}s$ e grade Tipo 1.

Os resultados para a grade do Tipo 2 para os modelos $\mathcal{M}_4$ e $\mathcal{M}_5$ são similares aos da grade do Tipo 1 e estão disponíveis no Apêndice A.

De maneira resumida, os melhores modelos são apresentados na Tabela 5.12. Nas situações em que os critérios não foram concordantes, optamos por escolher os melhores modelos pelo valor de $LPML$ e/ou $WAIC$, uma vez que o DIC pode apresentar problemas (Spiegelhalter et al., 2014). Ratificando, $\mathcal{M}_{4_{\phi=0.7}}$ foi o modelo que melhor se ajustou aos dados de câncer de mama, quando comparado aos demais, apresentando os menores valores de $LPML, WAIC$ e DIC. Observamos também através dos critérios, que há uma maior quantidade de modelos indicando usar a grade do Tipo 1 para os mais diferentes valores de $m^{'}$ e que somente os modelos $\mathcal{M}_0$, $\mathcal{M}_1$ e $\mathcal{M}_2$ indicam utilizar uma grade o Tipo 2 com tamanhos grandes de $m^{'}$ iguais a 41, 41 e 30, respectivamente.

Tabela 5.12: Modelos com menor $LPML$, DIC e $WAIC$.

Modelos	$m^{'}$	$LPML$	$WAIC$	DIC	Tipo da grade
$\mathcal{M}_0$	41	304,74	302,76	297,80	2
$\mathcal{M}_1$	41	299,72	299,01	292,22	2
$\mathcal{M}_2$	30	289,72	-	288,92	2
$\mathcal{M}_3$	10	304,96	304,43	314,74	1
$\mathcal{M}_{4_{\phi=0.7}}$	20	286,92	286,85	289,56	1
$\mathcal{M}_{5_{\phi=0.6}}$	20	291,25	291,14	296,91	1

As estimativas dos parâmetros de interesse dos melhores modelos com β fixo no tempo, $\mathcal{M}_0$ e $\mathcal{M}_3$, respectivamente, são apresentadas na Tabela 5.13. Note que as estimativas para β representam o efeito de grupo (quimioterapia e radioterapia) e os valores de $\exp\{\beta\}$ representam as estimativas da razão de riscos entre um paciente do grupo quimioterapia com um do grupo quimioterapia e radioterapia. No modelo $\mathcal{M}_0$ o efeito do grupo tratamento é não significativo, pois o intervalo HPD de 95% contém o valor 0. A significância da covariável tratamento é observada no modelo $\mathcal{M}_3$. No entanto, quando observamos o intervalo HPD $\exp\{\beta\}$ percebemos que o tratamento combinado é não significativo em ambos os modelos.

Tabela 5.13: Resumos *a posteriori* para os modelos selecionados.

Modelos	Parâmtros	Média	Intervalo HPD 95%	
$\mathcal{M}_0$	β	0,52	-0,02	1,07
	$\exp\{\beta\}$	1,75	0,86	2,72
$\mathcal{M}_3$	β	1,03	0,14	2,11
	$\exp\{\beta\}$	3,20	0,48	6,46

Uma forma de comparar os melhores modelos ajustados consiste confrontar a função de sobrevivência empírica com a ajustada pelo modelo. No caso de censura à direita, comparar a curva de sobrevivência estimada com o estimador de Kaplan-Meier (Kaplan & Meier, 1958) é uma forma interessante de saber se o modelo em questão está bem ajustado. Para dados com censura intervalar uma extensão ao estimador de Kaplan-Meier é o estimador não paramétrico proposto por Turnbull (1976), que será utilizado para avaliar a adequação dos modelos ajustados. Dessa forma, apresentamos as funções de sobrevivência estimadas nas Figuras 5.16 a 5.21, em comparação com a função de sobrevivência empírica estimada via estimador de Turnbull. Observamos que o modelo $\mathcal{M}_0$ (Figura 5.16) foi o modelo que apresentou a pior adequação aos dados, por não acompanhar a estimativa de Turnbull. No outro extremo, os modelos $\mathcal{M}_2$ e $\mathcal{M}_4$ (Figuras 5.19 e 5.20) apresentaram visualmente a melhor adequação aos dados de câncer de mama assim como os critérios haviam confirmado. Todos os ajustes e comparações podem encontrados nas figuras a seguir.

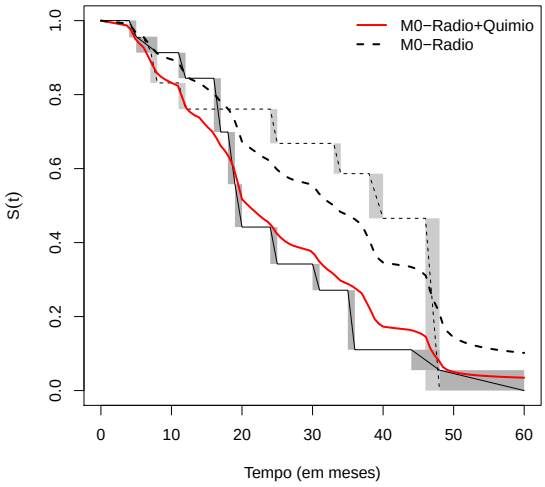

Figura 5.16: Estimativa da função de sobrevivência para o modelo $\mathcal{M}_0$

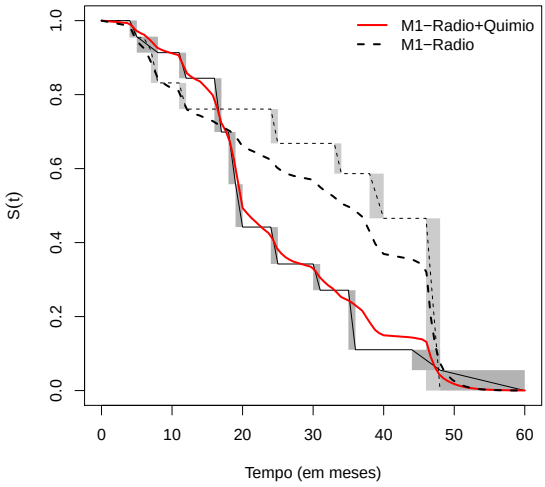

Figura 5.17: Estimativa da função de sobrevivência para o modelo $\mathcal{M}_1$

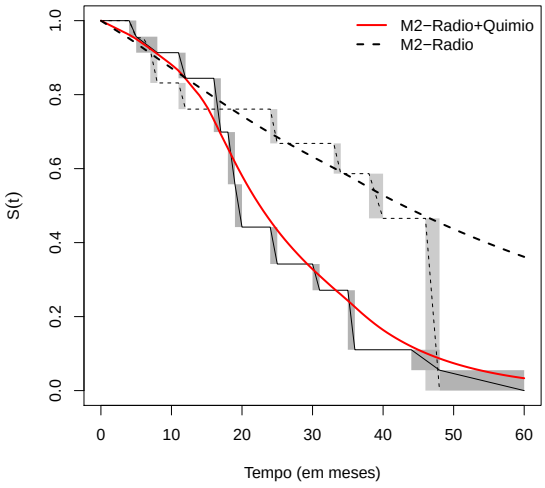

Figura 5.18: Estimativa da função de sobrevivência para o modelo $\mathcal{M}_2$

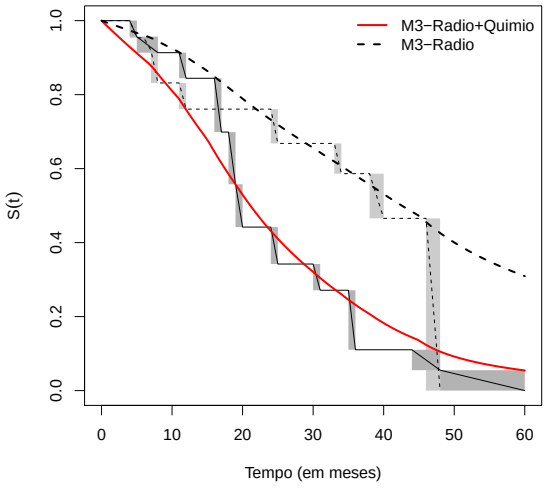

Figura 5.19: Estimativa da função de sobrevivência para o modelo $\mathcal{M}_3$

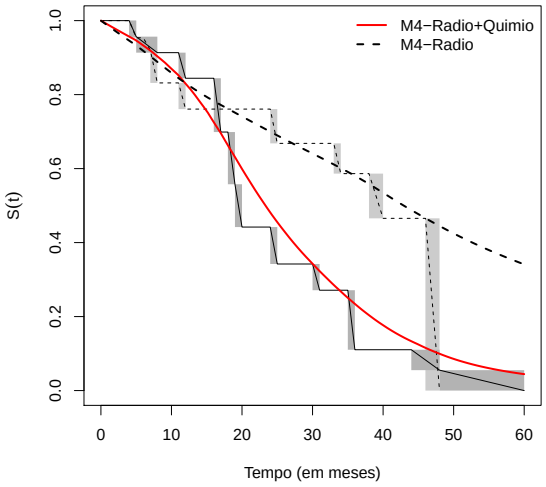

Figura 5.20: Estimativa da função de sobrevivência para o modelo $\mathcal{M}_4$

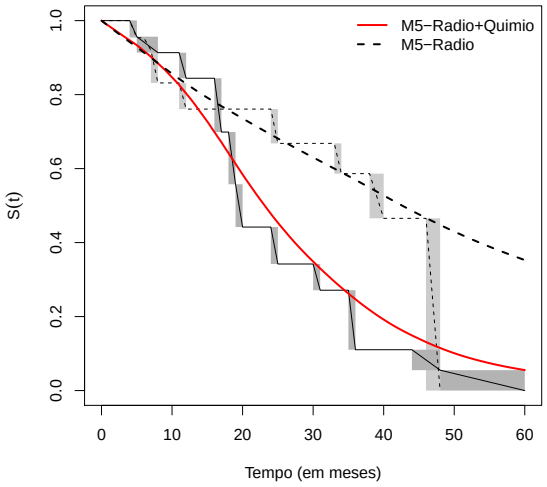

Figura 5.21: Estimativa da função de sobrevivência para o modelo $\mathcal{M}_5$

As Figuras 5.22 a 5.25 apresentam a estimativa do efeito do tratamento ao longo do tempo para os modelos $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_4$ e $\mathcal{M}_5$. Podemos notar que o modelo $\mathcal{M}_1$ (Figuras 5.22) é o modelo que apresenta maior variabilidade na estimativa de seu coeficiente ao longo do tempo, quando comparado com os demais modelos. Uma possível justificativa para tal resultado pode ser o uso de distribuições gamas independentes para as taxas do MEP, pela pouca presença de informações por intervalo. Por outro lado, as curvas mais suaves, ou seja, de menor variabilidade, são referentes aos modelos $\mathcal{M}_4$ e $\mathcal{M}_5$ (Figuras 5.24 e 5.25). Observe também, que em todas as figuras há um aumento do efeito da covariável tratamento em torno de 20 semanas.

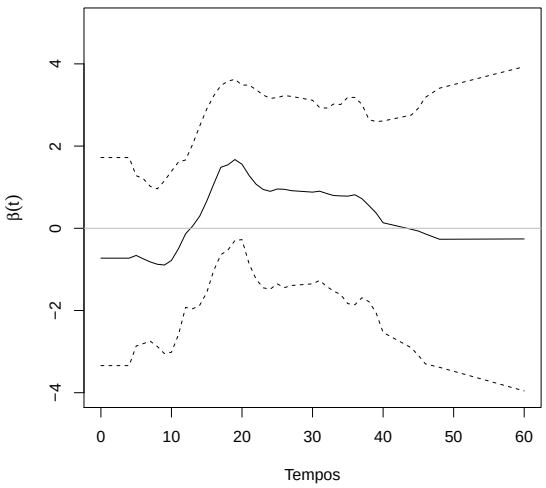

Figura 5.22: Estimativa do efeito da covariável radioterapia e químioterapia ao longo do tempo para o modelo $\mathcal{M}_1$. (As linhas tracejadas representam os intervalos de credibilidade de 95%.)

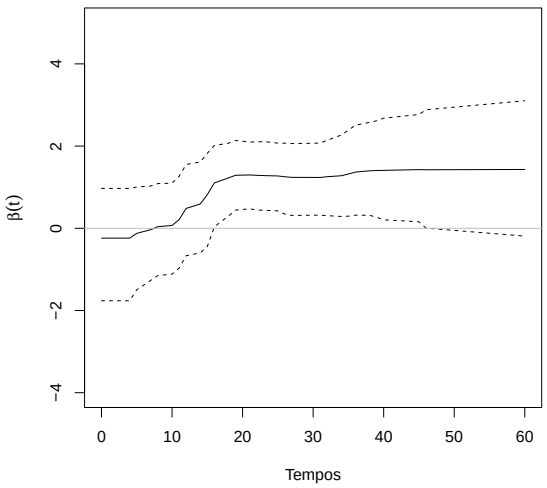

Figura 5.23: Estimativa do efeito da covariável radioterapia e químioterapia longo do tempo para o modelo $\mathcal{M}_2$. (As linhas tracejadas representam os intervalos de credibilidade de 95%.)

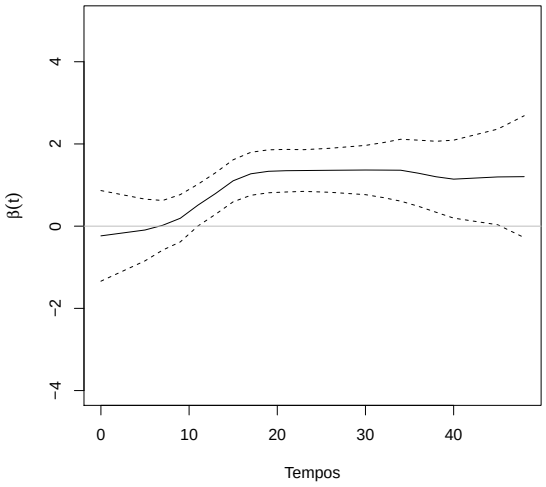

Figura 5.24: Estimativa do efeito da covariável radioterapia e químioterapia longo do tempo para o modelo $\mathcal{M}_4$. (As linhas tracejadas representam os intervalos de credibilidade de 95%.)

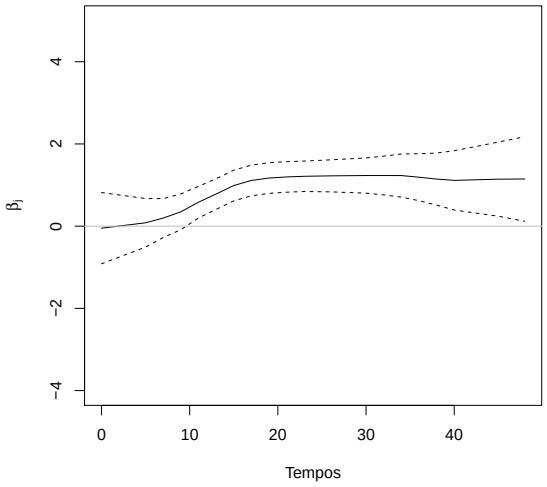

Figura 5.25: Estimativa do efeito da covariável radioterapia e químioterapia longo do tempo para o modelo $\mathcal{M}_5$. (As linhas tracejadas representam os intervalos de credibilidade de 95%.)

Na Tabela 5.14 apresentamos alguns resultados descritivos sobre a modelagem da grade envolvendo os modelos $\mathcal{M}_2$, $\mathcal{M}_3$ e $\mathcal{M}_5$. O modelo $\mathcal{M}_2$ mostrou que, iniciando $m' = 30$, o número mais provável de intervalos é em média 6,49 com intervalo HPD de 95% igual a [2;13]. Para os modelos baseados no MPP, observamos $\mathcal{M}_3$, que a moda *a posteriori* é da ordem de 3 intervalos na grade com HPD de 95% igual a [2;5]. A partição mais provável *a posteriori* neste igual a $\{0, 7, 15, 23, \infty\}$ com probabilidade 0.0716 (com valor de $m' = 10$). Para o modelo $\mathcal{M}_{5_{\phi=0,6}}$, partindo de $m' = 20$, temos moda *a posteriori* igual a 11 intervalos e HPD de 95% igual a [6;14]; neste caso, a partição mais provável é igual a $\{0, 5, 11, 15, 21, 26, 38, 40, 45, 48, \infty\}$ com probabilidade 0,002.

Tabela 5.14: Estatísticas descritivas para o número de intervalos *a posteriori*.

Modelos	Mínimo	Máximo	Média	Mediana	Moda	Desvio-padrão	HPD 95%
$\mathcal{M}_2$	1	21	6,49	6	5	3.22	[2;13]
$\mathcal{M}_3$	1	8	3,28	3	3	1,07	[2;5]
$\mathcal{M}_{5_{\phi=0,6}}$	4	19	10,53	11	11	2.20	[6;14]

5.2 Análise de dados com fração de cura

Nesta seção apresentamos os resultados dos modelos aplicados aos dados simulados e dados reais com fração de curados. Para esta análise, utilizamos a seguinte nomenclatura para os modelos:

$\mathcal{M}_0$: Modelo de fração de cura dinâmico com grade fixa;

$\mathcal{M}_1$: Modelo de fração de cura dinâmico com grade aleatória.

5.2.1 Reescrevendo a partição mais fina

Os modelos desenvolvidos nesta tese utilizam o algoritmo de ampliação de dados para o ajuste dos modelos. Devido a fração de curados, uma região do platô é formada e não irá apresentar muitos indivíduos com tempo de falha e dessa forma, os resultados dos modelos podem apresentar problemas de instabilidade numérica nas estimativas das taxas do MEP induzidas por $\tau^{'}$. Assim, para o ajuste dos modelos com fração de cura, definimos um limiar para o início da região do platô. Por exemplo, na Figura 5.26, apresentamos a estimativa da função de sobrevivência, via estimador de Turnbull, para os dados de pacientes hemofílicos. Note que, a partir de aproximadamente 26 meses, é formado um platô nos indicando a presença de poucos tempos de infecção após este limiar.

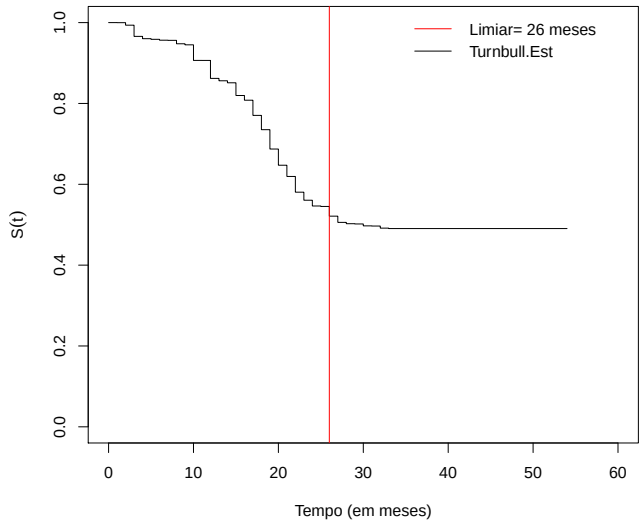

Figura 5.26: Estimativa da função de sobrevivência para pacientes hemofílicos, via estimador Turnbull (Turnbull, 1976).

Através do limiar do platô, estipulamos um certo percentual de intervalos na região do platô, como uma solução para o problema. Assim, é permitido determinar uma quantidade menor de intervalos nesta região através do limiar.

A quantidade de intervalos após limiar de tempo é obtida através de um percentual baseado no número máximo $m^{'}$ de intervalos admitidos *a priori*. Por exemplo, se $m^{'} = 10$ e

se estipulamos um percentual de 10% de intervalos após o limiar do platô, teremos somente um intervalo após o limiar determinado.

5.2.2 Dados simulados

Para gerarmos dados com fração de cura a partir do modelo de tempos de promoção, assumimos que o número de causas latentes M segue uma distribuição Poisson, em que $M \sim Poisson(\theta)$, com $\theta = \exp\left\{\boldsymbol{z}^\top \boldsymbol{\psi}\right\}$. Para a geração dos tempos de promoção, tomamos o mínimo de M tempos de promoção obtidos a partir de uma distribuição de Weibull, com a seguinte função densidade

$$f(r) = \alpha\gamma r^{\alpha-1} \exp\left\{-\gamma r^{\alpha}\right\}, \quad r > 0, \alpha > 0, \gamma > 0.$$

O algoritmo para a geração dos dados de sobrevivência com censura intervalar para o modelo de tempos de promoção é o seguinte:

1. Para $i = 1, \ldots, n$, geramos M_i através de uma distribuição Poisson com média $\exp\{\psi_0 + \psi_1 z_i\}$, em que $z_i \sim$ Bernoulli(0.5).

2. Se $M_i = 0$, assuma que $T_i = \infty$ representando um indivíduo curado, caso contrário, gere M_i tempos de promoção de uma distribuição Weibull com parâmetros α e γ e tome o mínimo entre os tempos gerados para representar o tempo do evento de interesse T_i.

3. Para $i = 1, \ldots, n$, gere os tempos de censura como $C_i = \min(a, b \times A)$, em que $A \sim \text{Exp}(\zeta)$, em que a e b são constantes relacionadas a proporção de censura, e tome $\delta_i = I(T_i \leq C_i)$.

4. Para a construção dos intervalos de tempo utilize os passos 3 e 4 descritos na Subseção 5.1.1.

Ao utilizar o algoritmo de geração dos dados, optamos em utilizar um tamanho amostral igual a 500. Os valores dos parâmetros para a geração dos dados foram os seguintes:

- No preditor linear, assumimos $\psi_0 = 0.5$ e $\psi_1 = -2.5$;
- Os parâmetros da distribuição Weibull foram, $\alpha = 1,5$ e $\gamma = 0,03$;
- Para determinar o percentual de censura e de curados utilizamos os seguintes valores de a e b iguais a 27 e 15, respectivamente, o que nos fornece um tempo de no máximo de 27 e $\zeta = 0,01$.
- Os valores de d_1 e d_2 foram $0,1$ e $0,5$, respectivamente.

As especificações *a priori* para os modelos $\mathcal{M}_0$ e $\mathcal{M}_1$ foram as seguintes: $\alpha_0 = 0.01$ e $\gamma_0 = 0.01$ e uma distribuição *a priori* $\phi \sim Unif(1, 1)$ para o fator de desconto. Para os

coeficientes de regressão assumimos uma distribuição $N(\mu, \sigma^2)$, em que $\psi_0 \sim N(0,1)$ e $\psi_1 \sim N(0,100)$,. A escolha da variância de ψ_0 é para garantir a estabilidade nas estimativas dos parâmetros de interesse dos modelos . Exceto para ψ_0, todos os valores especificados *a priori* uma informação vaga para os parâmetros de interesse. Comparamos os modelos $\mathcal{M}_0$ e $\mathcal{M}_1$ usando a grade inicial somente do Tipo 1 nos valores de m' iguais a 10, 20, 30 e 40. Para a realização da análise bayesiana foi considerado um total de 100000 iterações, com um aquecimento da cadeia de 50000 iterações, um espaçamento de 10, resultando em uma amostra *a posteriori* de tamanho 5000.

Na Tabela 5.15 apresentamos dos valores dos critérios de seleção para os modelos $\mathcal{M}_0$ e $\mathcal{M}_1$. De acordo com o critério $LPML$, o modelo que melhor se ajustou aos dados foi o modelo $\mathcal{M}_1$ com $m' = 20$. Por outro lado, o $WAIC$ nos mostra que o melhor também é o modelo $\mathcal{M}_1$ com $m' = 10$. E finalmente, o critério DIC nos indica que o modelo que melhor se ajusta aos dado é o modelo $\mathcal{M}_1$ com grade inicial contendo 40 intervalos. Por simplicidade, optamos pelo modelo mais parcimonioso. Dessa forma, baseamos toda a análise através dos ajustes obtidos pelos modelos $\mathcal{M}_0$ e $\mathcal{M}_1$, com $m' = 10$ e grade do Tipo 1.

Tabela 5.15: Critérios $LPML$, $WAIC$ e DIC para os modelos $\mathcal{M}_0$ e $\mathcal{M}_1$.

Critérios	Modelos	m'			
		10	20	30	40
$LPML$	$\mathcal{M}_0$	2737,80	2737,23	2737,45	2736,45
	$\mathcal{M}_1$	2735,71	2735,64	2736,96	2737,18
$WAIC$	$\mathcal{M}_0$	2742,31	2741,75	2741,93	2740,89
	$\mathcal{M}_1$	2740,06	2740,11	2741,36	2741,72
DIC	$\mathcal{M}_0$	3608,25	3608,05	3618,62	3607,68
	$\mathcal{M}_1$	3602,37	3628,99	3590,45	3623,23

São apresentados nas Figuras 5.27, 5.28 e 5.29 os valores dos critérios de seleção de acordo com o valor de m' e o tipo de grade. Observamos que pelas Figuras 5.27 e 5.28, para a grade do Tipo 1, que no para o modelo $\mathcal{M}_1$, que a medida que o tamanho da grade inicial aumenta a qualidade de ajuste do modelo aos dados diminui.

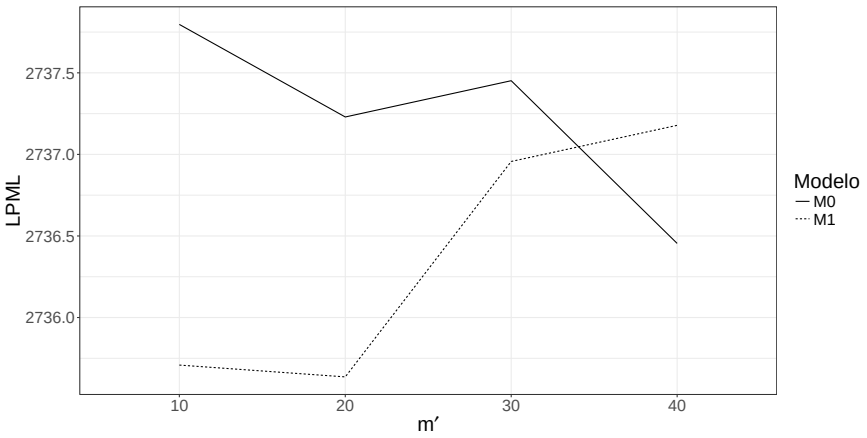

Figura 5.27: Comparação dos modelos $\mathcal{M}_0$ e $\mathcal{M}_1$ via $LPML$.

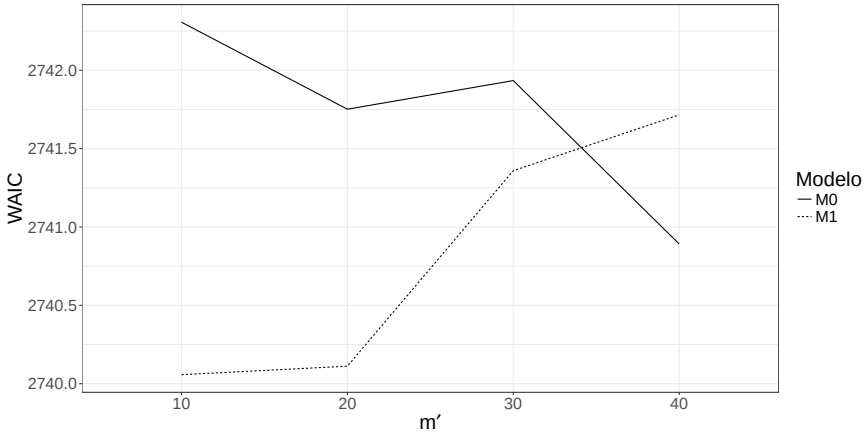

Figura 5.28: Comparação dos modelos $\mathcal{M}_0$ e $\mathcal{M}_1$ via $WAIC$.

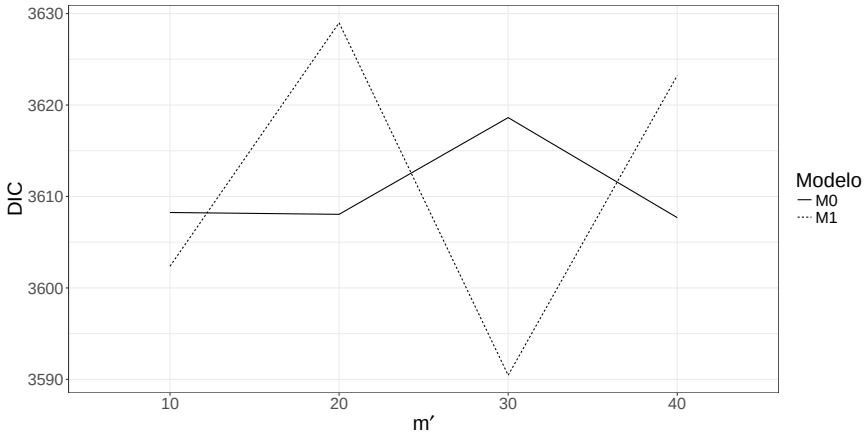

Figura 5.29: Comparação dos modelos $\mathcal{M}_0$ e $\mathcal{M}_1$ via DIC.

Na Figura 5.30 apresentamos as amostras *a posteriori* para o fator de desconto, para os modelos $\mathcal{M}_0$ e $\mathcal{M}_1$, grade do Tipo 1, e para as diferentes escolhas de $m^{'}$. Os resultados indicam que, independente do tipo de grade, a distribuição *a posteriori* do fator de desconto associada ao modelo $\mathcal{M}_0$ está sempre concentrada em maiores valores que aqueles associados ao modelo $\mathcal{M}_1$. Uma possível explicação para isso é que $\mathcal{M}_0$ contém mais intervalos que $\mathcal{M}_1$ e assim, pelo maior número de intervalos, menor informação ficará disponível em cada intervalo, compensando assim com o aumento do fator de desconto.

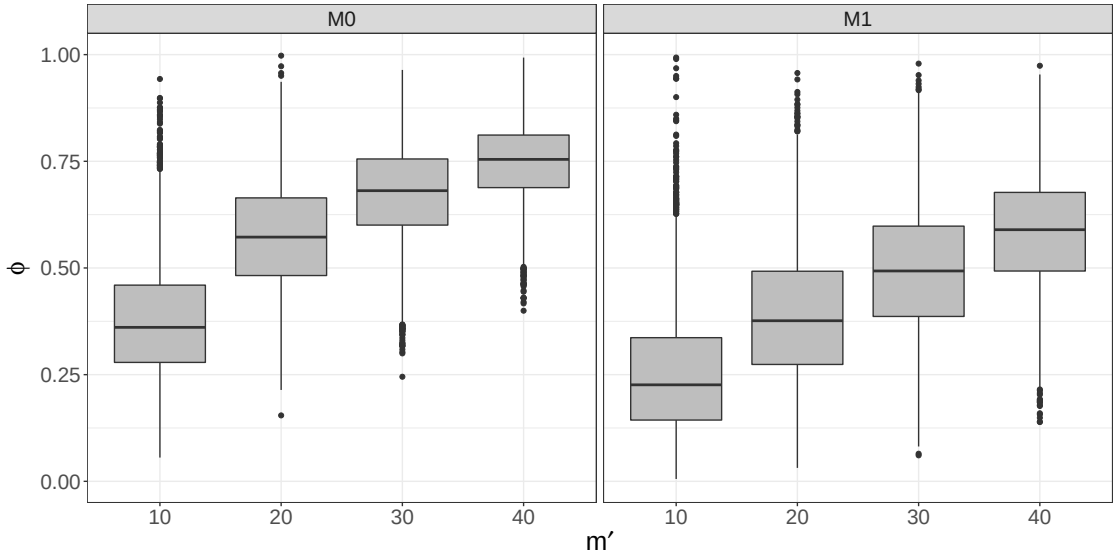

Figura 5.30: Amostra *a posteriori* para ϕ no modelo $\mathcal{M}_0$ e $\mathcal{M}_0$, com grade do Tipo 1.

A seguir apresentamos as amostras *a posteriori* para os coeficientes da regressão conjuntamente com o valor real atribuído na geração dos dados, com os seus respectivos intervalos HPD de 95%. Para o modelo $\mathcal{M}_0$ observamos nas Figuras 5.31 e 5.32 as amostras de β_0 e β_1 nos indicam que a cadeia está convergindo para os valores reais, 0,45 e -2,50, respectivamente. Observamos também, que para os diferentes tipos valores de $m^{'}$ os limites dos intervalos HPD contemplam 95% incluem os valores reais.

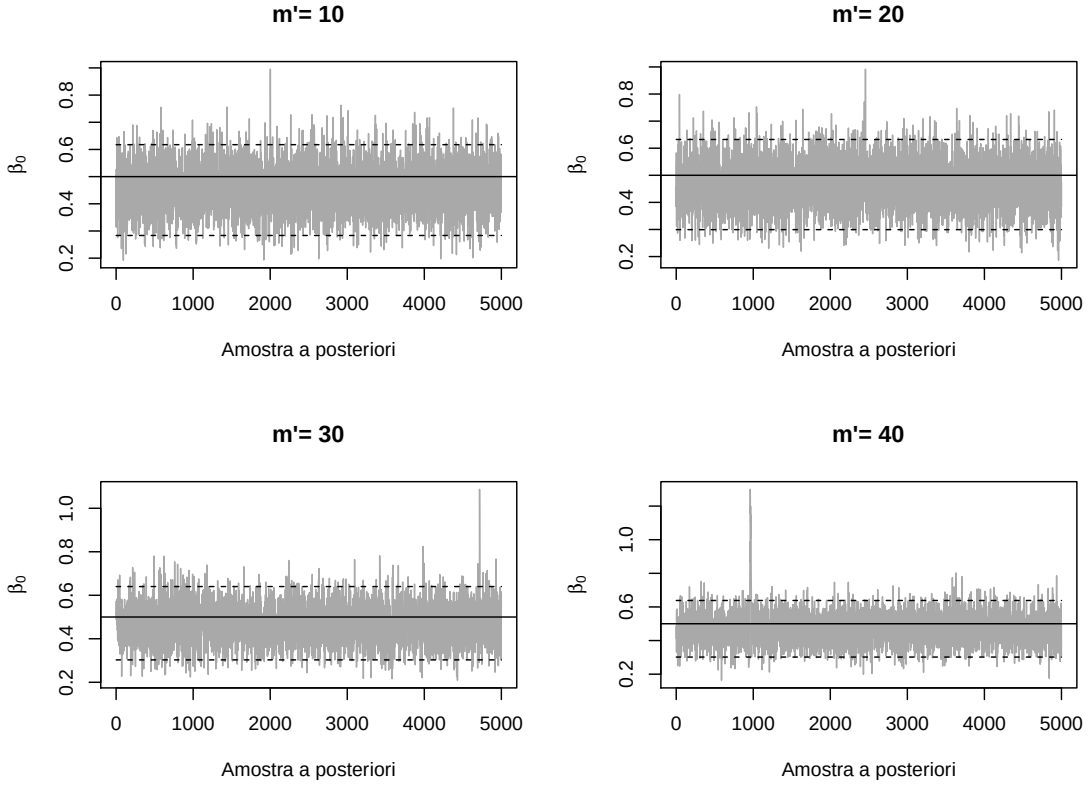

Figura 5.31: Amostra *a posteriori* para β_0 no modelo $\mathcal{M}_0$ e grade do Tipo 1 (Valor real na linha escura).

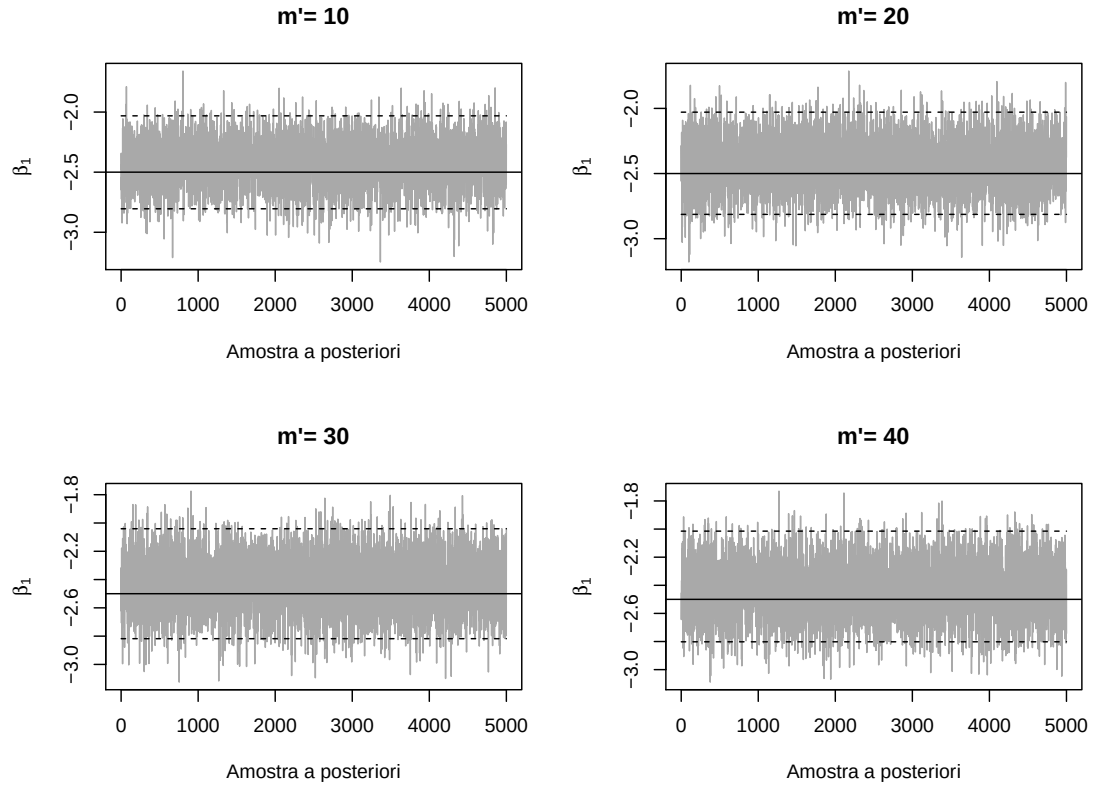

Figura 5.32: Amostra *a posteriori* para β_1 no modelo $\mathcal{M}_0$ e grade do Tipo 1 (Valor real na linha escura).

Com relação ao modelo $\mathcal{M}_1$, apresentamos na Figuras 5.33 e 5.33 as amostras *a posteriori* para β_0 e β_1, respectivamente. Neste caso, os resultados foram simulares os obtidos pelo modelo $\mathcal{M}_0$. Para a grade do Tipo 2, os resultados foram simulares e encontram-se no Apêndice A

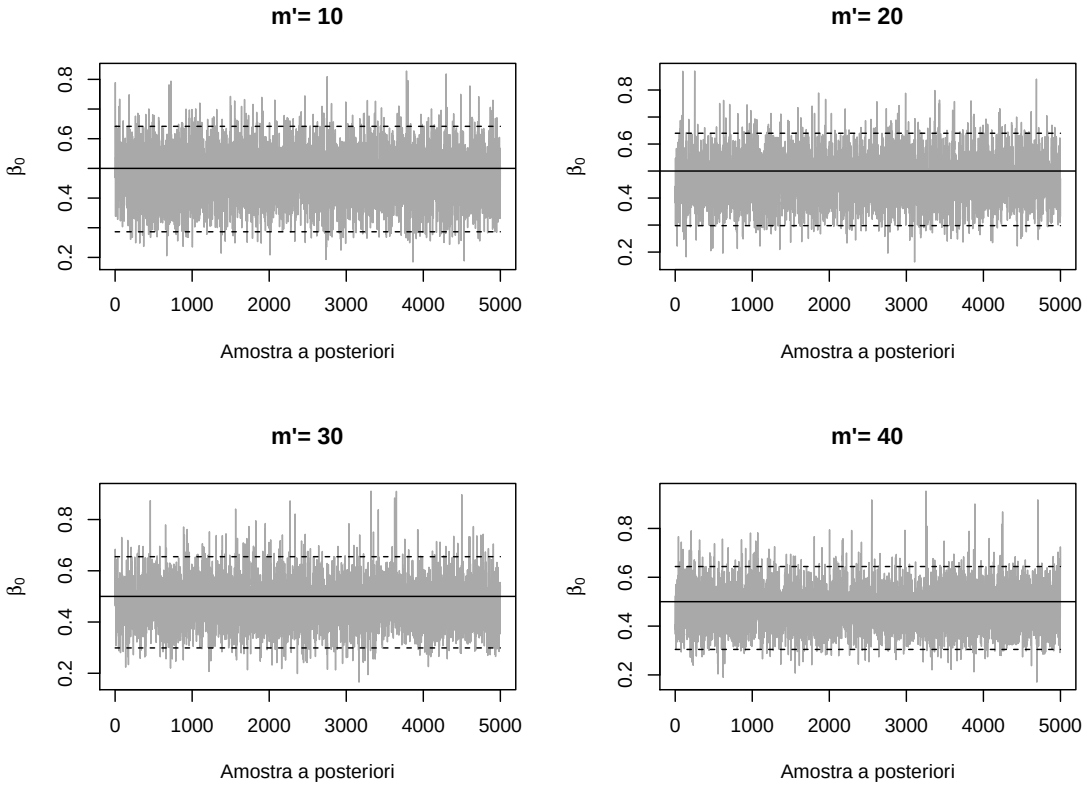

Figura 5.33: Amostra *a posteriori* para β_0 no modelo $\mathcal{M}_1$ e grade do Tipo 1 (Valor real na linha escura).

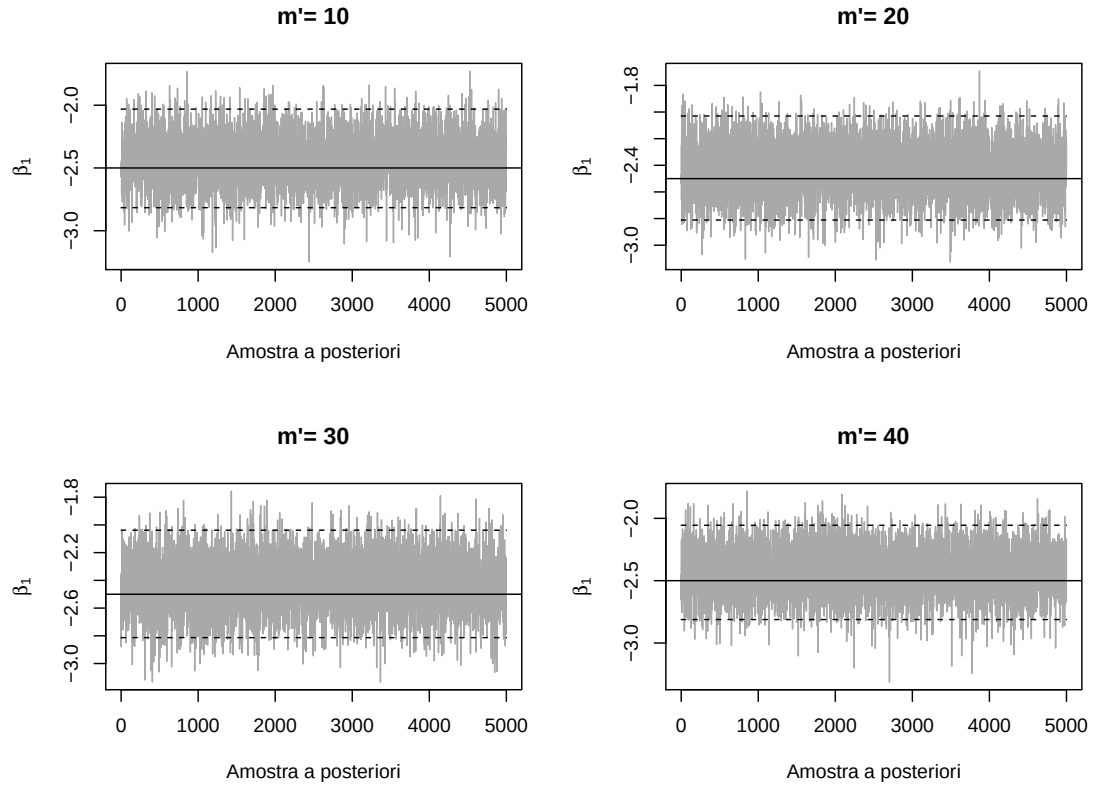

Figura 5.34: Amostra *a posteriori* para β_1 no modelo $\mathcal{M}_1$ e grade do Tipo 1 (Valor real na linha escura).

Apresentamos as funções de risco basal estimada para os modelos $\mathcal{M}_0$ e $\mathcal{M}_1$, respectivamente. Observamos em todas as comparações dos modelos e para os diferentes valores de $m^{'}$, um comportamento côncavo assim como estipulado na geração dos dados, no entanto, os modelos com $m^{'} = 10$ foi o modelo que mais se aproximou mais do comportamento real.

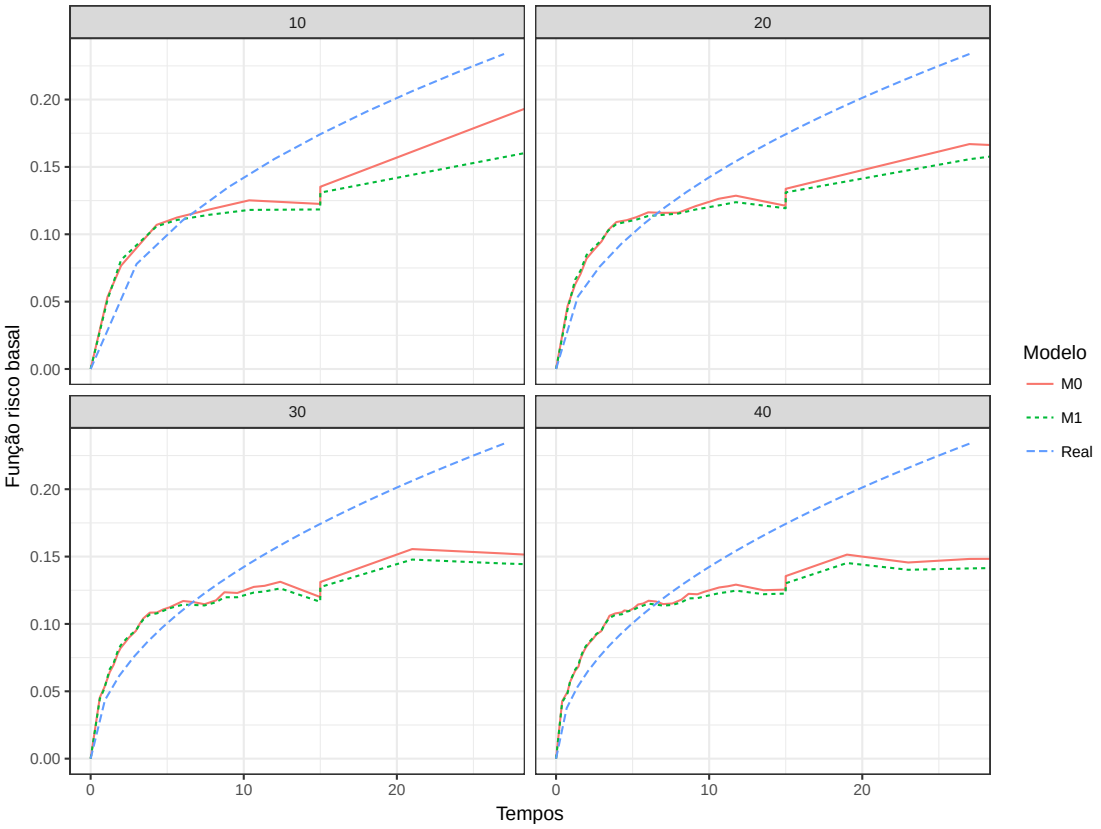

Figura 5.35: Estimativa *a posteriori* para a função risco basal para os tempos de promoção.

As funções de sobrevivência populacionais estimadas pelos modelos $\mathcal{M}_0$ e $\mathcal{M}_1$ são apresentadas nas Figuras 5.36 e 5.37, respectivamente, conjuntamente com as funções de sobrevivências reais e via estimador de Turnbull. Observamos que para ambos modelos, as estimativas das funções de sobrevivência populacional acompanham de maneira satisfatória a curva de sobrevivência real, por outro lado notamos que tanto o modelos $\mathcal{M}_0$ e $\mathcal{M}_1$ acompanham bem o comportamento do estimador de Turnbull.

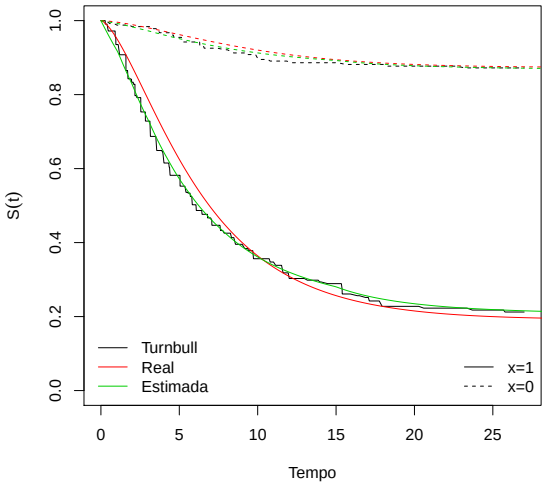

Figura 5.36: Estimativa da função de sobrevivência populacional para o modelo $\mathcal{M}_0$ (linha em vermelho) e curvas reais (linha em preto).

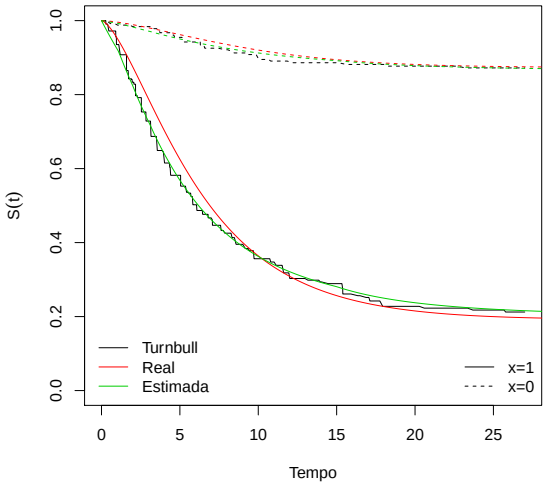

Figura 5.37: Estimativa da função de sobrevivência populacional para o modelo $\mathcal{M}_1$ (linha em vermelho) e curvas reais (linha em preto).

O valor real da fração de cura para $x = 1$ é dada por

$$\exp\left(-\exp\left(0,5-2,5\right)\right) = 0,87.$$

Assim, apresentamos na Tabela 5.16 as estimativas pontuais e intervalares para a fração de cura quando $x = 1$. Observamos que todas as estimativas pontuais estão próximas do valor real, para os diferentes valores de m', e todos os intervalos HPD de 95% contém o verdadeiro valor.

Tabela 5.16: Estimativas da fração de cura para os modelos $\mathcal{M}_0$ e $\mathcal{M}_1$ para $x = 1$.

Modelo	m'	Real	Média	Mediana	Moda	Desvio padrão	HPD 95%	
$\mathcal{M}_0$	10	0,87	0,8678	0,8690	0,8970	0,0226	0,8211	0,9083
	20		0,8663	0,8680	0,8925	0,0236	0,8222	0,9136
	30		0,8660	0,8672	0,8855	0,0236	0,8179	0,9081
	40		0,8658	0,8668	0,8869	0,0240	0,8205	0,9124
$\mathcal{M}_1$	10	0,87	0,8657	0,8668	0,8681	0,0233	0,8194	0,9105
	20		0,8658	0,8670	0,8673	0,0230	0,8230	0,9111
	30		0,8659	0,8673	0,8686	0,0232	0,8227	0,9126
	40		0,8650	0,8661	0,8786	0,0228	0,8194	0,9075

Quando $x = 0$, temos que o valor real da fração de cura dada por

$$\exp\left(-\exp\left(0,5\right)\right) = 0,19.$$

Na Tabela 5.17 as estimativas pontuais e intervalares para a fração de cura quando $x = 0$. Assim como para o caso anterior, notamos que todas as estimativas pontuais estão próximas do valor real, como também, os intervalos HPD de 95% contém o verdadeiro valor, para os diferentes valores de $m^{'}$.

Tabela 5.17: Estimativas da fração de cura para os modelos $\mathcal{M}_0$ e $\mathcal{M}_1$ para $x = 0$.

Modelo	$m^{'}$	Real	Média	Mediana	Moda	Desvio padrão	HPD 95%	
$\mathcal{M}_0$	10	0,19	0,2022	0,2021	0,1926	0,0286	0,1451	0,2586
	20		0,2020	0,2017	0,2589	0,0283	0,1501	0,2600
	30		0,2005	0,2003	0,1910	0,0289	0,1448	0,2585
	40		0,2002	0,2006	0,2231	0,0284	0,1481	0,2568
$\mathcal{M}_1$	10	0,19	0,2059	0,2054	0,2157	0,0277	0,1519	0,2598
	20		0,2033	0,2028	0,2296	0,0274	0,1520	0,2592
	30		0,2022	0,2021	0,1877	0,0275	0,1501	0,2583
	40		0,2019	0,2021	0,2470	0,0286	0,1503	0,2584

Observamos na Tabela 5.18 as estimativas descritivas da amostra *a posteriori* do número de intervalos referente ao modelo $\mathcal{M}_1$ com $m^{'} = 10$. Notamos que, a moda *a posteriori* foi igual a 6 intervalos com intervalo HPD 95% igual a $[3; 8]$. Podemos perceber também que, o número de intervalos *a posteriori* estabiliza, em geral, para um valor um pouco maior que a metade do número de intervalos admitidos na grade inicial ($m^{'} = 10$).

Tabela 5.18: Resumos descritivos para a amostra *a posteriori* do número de intervalos para o modelo $\mathcal{M}_1$.

Mínimo	Máximo	Média	Moda	Mediana	Desvio-padrão	HPD 95%
2	10	5,88	6	6	1,34	[3;8]

5.2.3 Dados de infecção por HIV-1 em pacientes hemofílicos

Os dados analisados nesta seção foram coletados como parte de um estudo prospectivo multicêntrico que teve por objetivo avaliar a taxa de infecção por HIV-1 entre pacientes com hemofilia (Goedert et al., 1989; Kroner et al., 1994). Em particular, os indivíduos do estudo formavam um grupo de risco para a contração de HIV-1 a partir de produtos derivados de sangue feitos a partir de doadores.

O estudo é caracterizado por se tratar de dados de sobrevivência com censura intervalar, pois não sabemos exatamente em qual momento os pacientes contraíram HIV-1. No entanto, temos o conhecimento dos intervalos de tempo (em meses) em que a soro conversão ocorreu, determinados pelos tempos de monitoramento.

Este estudo contou com um total de 544 pacientes, com um percentual de infecção de aproximadamente 49%. Com base na sua dose média anual de produtos derivados do sangue, os pacientes foram classificados em quatro grupos: alta, média, baixa ou nenhuma dose.

Neste trabalho, utilizamos os dados supracitados para ilustrar o modelo proposto em na Seção 4.2. Como variável preditora, dicotomizamos a variável do estudo original, para simplificar a análise, em 308 pacientes tratados com pelo menos alguma dose ($x = 0$) e 236 paciente tratados com nenhuma dose ($x = 1$). Os dados estão disponíveis no pacote `ICsurv` (McMahan & Wang, 2014).

As especificações *a priori* para os modelos $\mathcal{M}_0$ e $\mathcal{M}_1$ são: $\alpha_0 = 0.01$ e $\gamma_0 = 0.01$ e uma distribuição *a priori* Beta(θ_1, θ_2) para o fator de desconto, com $\theta_1 = 1, \theta_2 = 1$, representando uma distribuição uniforme no intervalo $[0, 1]$. Assumimos também, uma distribuição $N(\mu, \sigma^2)$, em que $\mu = 0$ e $\sigma^2 = 100$, para cada componente do vetor $\boldsymbol{\psi}$. A análise Bayesiana contou com um aquecimento das cadeias de 50000 iterações, espaçamento de 20 e um total de amostras *a posteriori* de 5000, contabilizando um total de 150000 iterações.

Baseado na grade mais fina τ' para limites dos distintos intervalos observados, grade do Tipo 1, observamos que o número de intervalos máximo é $m' = 28$. No entanto, para uma análise de sensibilidade, valores de m' iguais a 10 e 20 também foram considerados. Para a grade do Tipo 2, grade com intervalos equidistante, avaliamos os seguintes valores de m': 10, 20, 28, 30 e 40.

Os valores dos critérios de seleção ($LPML$, $WAIC$ e DIC) calculados para os dois modelos ajustados são apresentados na Tabela 5.19. De acordo com esses critérios, o modelo dinâmico com grade aleatória ($\mathcal{M}_1$) apresenta, de maneira geral, melhor ajuste quando comparado ao modelo $\mathcal{M}_0$, ou seja, utilizar a estrutura de agrupamento do MPP para modelar a grade dos tempos do MEP melhorou na qualidade de ajuste do modelo aos dados.

Tabela 5.19: Critérios de seleção para os modelos $\mathcal{M}_0$ e $\mathcal{M}_1$.

Critérios	Modelos	Tipo 1			Tipo 2				
		10	20	28	10	20	28	30	40
$LPML$	$\mathcal{M}_0$	1720,35	1733,09	1719,87	1732,30	1712,10	1745,08	1771,66	1758,37
	$\mathcal{M}_1$	1714,69	1735,40	1713,41	1729,31	1748,78	1727,48	1714,12	1693,04
$WAIC$	$\mathcal{M}_0$	1720,14	1731,88	1719,49	1731,54	1711,98	1744,35	1771,66	1757,98
	$\mathcal{M}_1$	1714,54	1734,53	1712,52	1728,90	1747,46	1726,98	1714,05	1692,57
DIC	$\mathcal{M}_0$	2592,21	2618,74	2566,50	2603,36	2537,00	2664,78	2791,71	2705,34
	$\mathcal{M}_1$	2552,82	2623,19	2570,79	2620,84	2685,45	2620,15	2542,32	2450,60

Uma comparação entre todos os modelos baseando-se no critério $LPML$ é apresentada na Figura 5.38. Para a grade do Tipo 2, observamos na Figura 5.38, que o modelo $\mathcal{M}_1$ apresenta melhor ajuste comparado ao modelo $\mathcal{M}_0$ à medida que o número de intervalos aumenta, a partir de $m^{'} = 28$.

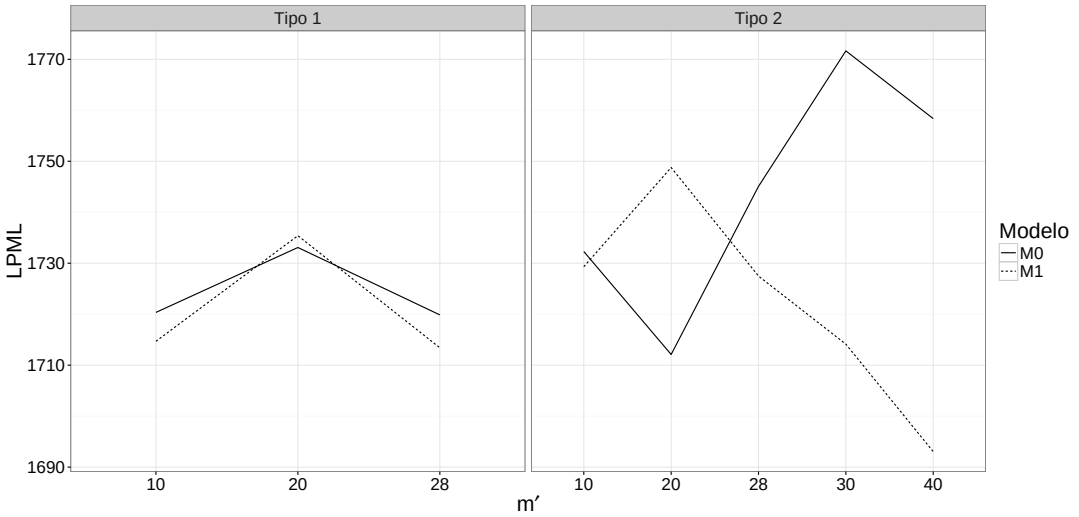

Figura 5.38: Valores de $LPML$ para os modelos ajustados.

Similarmente para o critério $LPML$, na Figura 5.39, os valores do critério $WAIC$ indicam que o modelo $\mathcal{M}_1$ fornece um melhor ajuste aos dados quando comparado ao modelo $\mathcal{M}_0$ para quase todos os valores de $m^{'}$ e tipos de grade. Observamos também que, assim como para o critério $LPML$, há uma perda de qualidade de ajuste no modelo $\mathcal{M}_1$ a medida que o valor de $m^{'}$ diminui.

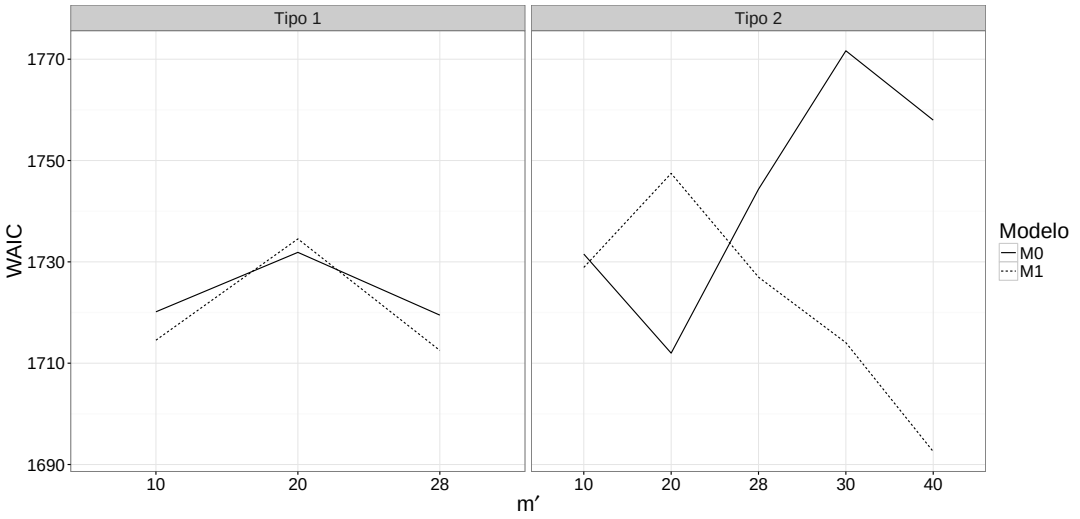

Figura 5.39: Valores de $WAIC$ para os modelos ajustados.

Na Figura 5.40 apresentamos os valores do critério DIC para os modelos em questão. Observamos que, na grade do Tipo 1, os modelos $\mathcal{M}_0$ e $\mathcal{M}_0$ apresentam resultados similares para os números de intervalos para $m^{'}$ igual à 20 e 28.

Um ponto importante a ressaltar é que, para todos os critérios, há um comportamento similar entre os modelos $\mathcal{M}_0$ e $\mathcal{M}_1$ para a grade do Tipo 1, diferentemente dos modelos ajustados utilizando a grade Tipo 2. No Tipo 2, para valores de $m^{'}$ maiores que 28 intervalos, há uma melhor qualidade de ajuste ao considerar mais intervalos.

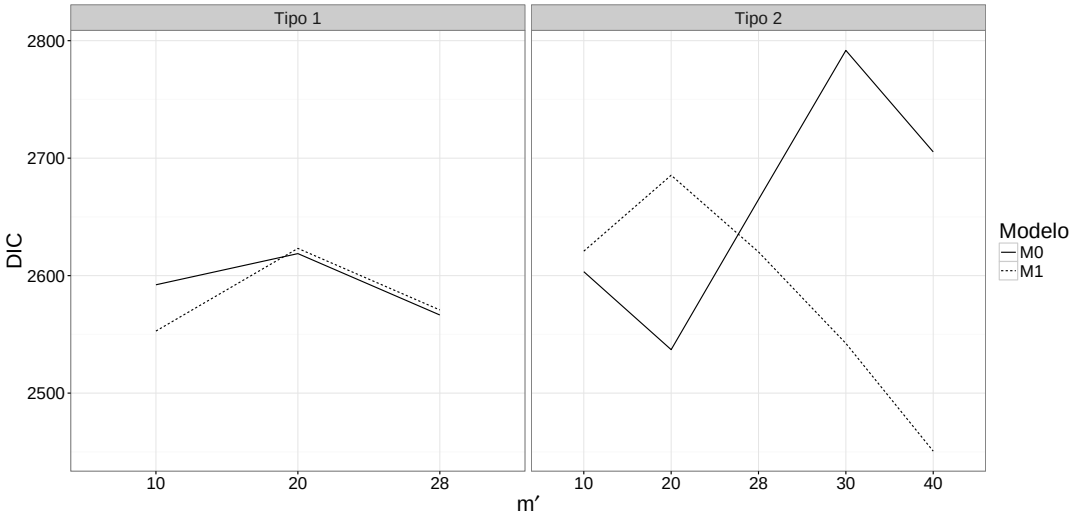

Figura 5.40: Valores de DIC para os modelos ajustados.

Em linhas gerais, com os valores de $LPML$ (1693,04), $WAIC$ (1692,57) e DIC (2450,60), respectivamente, para grade do Tipo 2, temos que o modelo $\mathcal{M}_1$ é o mais indicado para modelar os dados de pacientes hemofílicos quando comparado ao modelo $\mathcal{M}_0$. Dessa forma, optamos em modelar os dados de pacientes hemofílicos com $m^{'} = 40$ e grade do Tipo 2. Para fins comparativos, essa configuração de modelo será utilizada para o modelo dinâmico com grade fixa para analisar os dados de tempos de soroconversão.

A seguir apresentamos pelas Figuras 5.41 e 5.42 os gráficos de caixa das amostras *a posteriori* para o fator de desconto nos modelos $\mathcal{M}_0$ e $\mathcal{M}_1$ para a grade do Tipo 1 e 2. Observamos um resultado muito similar ao encontrado para os dados simulados.

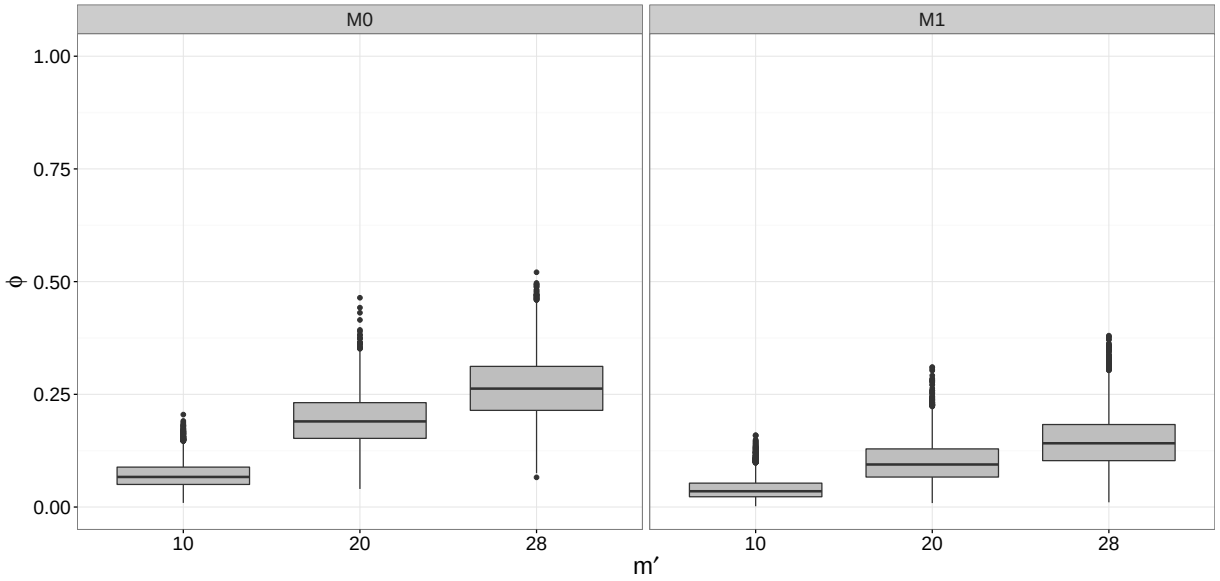

Figura 5.41: Gráficos de caixa das amostras *a posteriori* para o fator de desconto nos modelos $\mathcal{M}_0$ e $\mathcal{M}_1$ para a grade do Tipo 1.

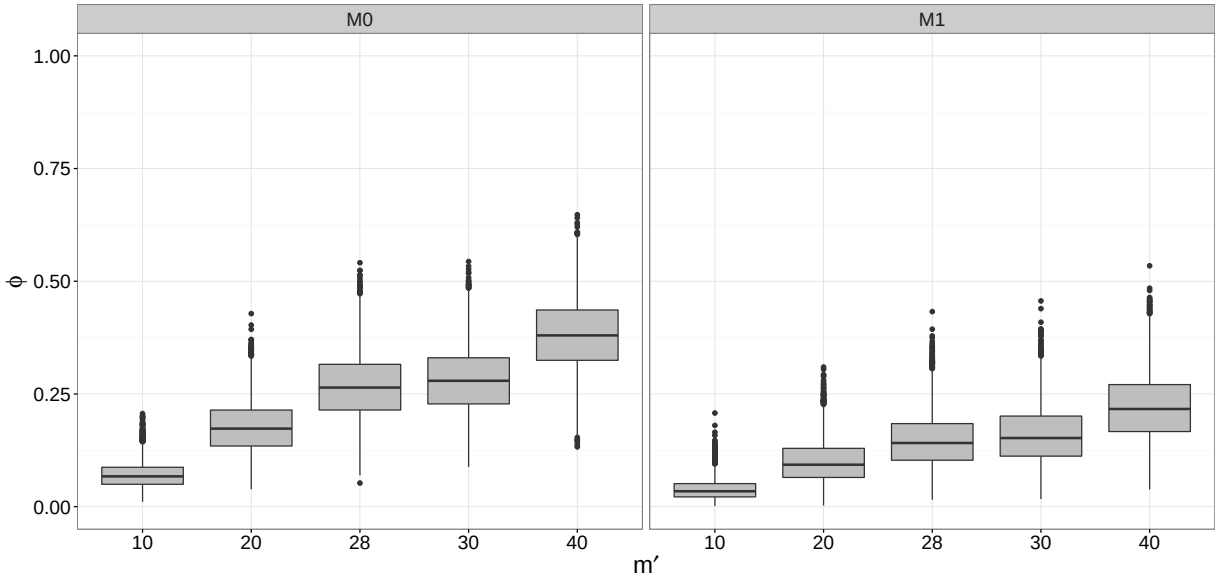

Figura 5.42: Gráficos de caixa das amostras *a posteriori* para o fator de desconto nos modelos $\mathcal{M}_0$ e $\mathcal{M}_1$ para a grade do Tipo 2.

As Figuras 5.43 e 5.44 apresentam as estimativas da função risco basal $h_0(t)$ para aos tempos de promoção. Observe que $h_0(t)$ apresenta um comportamento crescente para todos os valores de $m^{'}$ e tipos de grade, sugerindo que o risco de infecção aumenta a partir de aproximadamente 20 meses de acompanhamento.

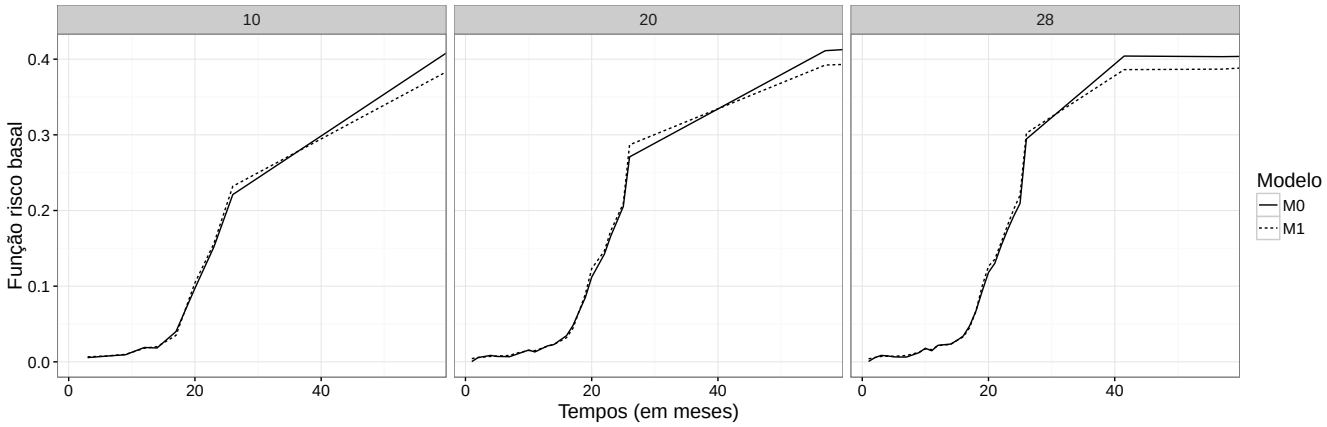

Figura 5.43: Estimativa *a posteriori* para a função risco basal para os tempos de promoção e grade do Tipo 1.

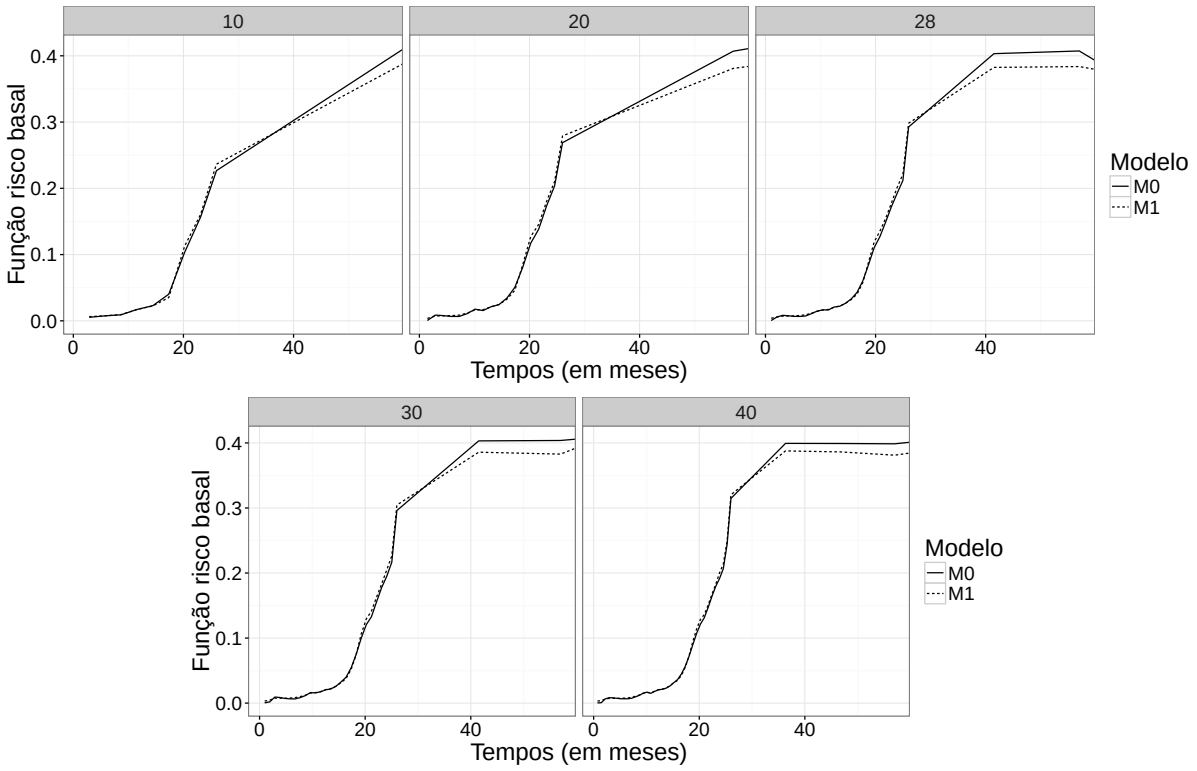

Figura 5.44: Estimativa *a posteriori* para a função risco basal para os tempos de promoção e grade do Tipo 2.

As estimativas dos parâmetros de interesse de cada modelos com as melhores qualidade de ajuste aos dados, modelo $\mathcal{M}_0$ e $\mathcal{M}_1$, com $m' = 40$ e grade Tipo 2, são apresentadas na Tabela 5.20. De maneira geral, ambos modelos ajustados mostram que a variável tratamento é significativa uma vez que o intervalos HPD 95% não contém o valor 0, ou seja, há diferença entre o grupo que não recebe nenhuma dose com o tratado com alguma dose. Em outras palavras, observamos que não receber dose aumenta a probabilidade dos pacientes hemofílicos de não contrair HIV-1.

Tabela 5.20: Resultados descritivos dos modelos mais bem ajustados segundo os critérios de seleção, com $m' = 40$ e grade do Tipo 2.

Modelos	Parâmetros	Média	Desvio padrão	HPD 95%
$\mathcal{M}_0$	ψ_0	0,43	0,07	[0,30;0,58]
	ψ_1	-2,49	0,20	[-2,88;-2,12]
	ϕ	0,22	0,07	[0,08;0,36]
$\mathcal{M}_1$	ψ_0	0,44	0,07	[0,31;0,58]
	ψ_1	-2,50	0,20	[-2,88;-2,10]
	ϕ	0,38	0,08	[0,23;0,53]

Podemos observar na Tabela 5.21 que o número mais provável de intervalos é igual a 22 (com probabilidade de 0,1272) e o intervalo HPD de 95% igual a [16;27]. É possível notar que neste ajuste, de um número de intervalos admitidos *a priori* igual a 40, o modelo reduziu para pouco mais da metade. Outras medidas podem ser encontradas na Tabela 5.21, assim como a distribuição do número de intervalos apresentado na Figura 5.45.

Tabela 5.21: Resumos descritivos para a distribuição *a posteriori* do número de intervalos para o modelo $\mathcal{M}_1$, com $m' = 40$.

Mínimo	Máximo	Média	Moda	Mediana	Desvio padrão	HPD 95%
8	33	21.28	22	21	3.06	[16;27]

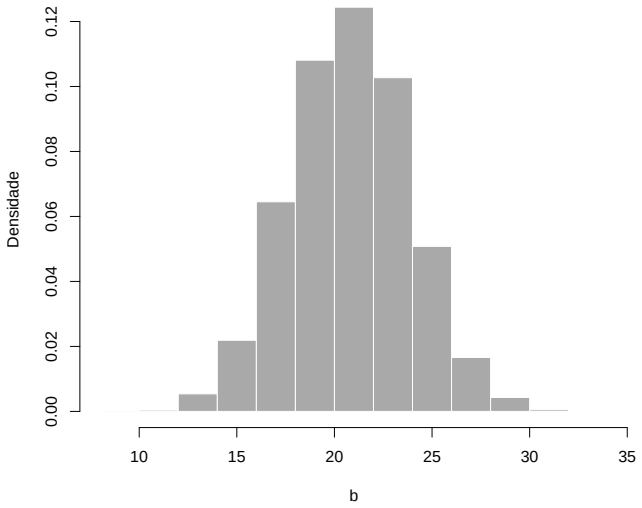

Figura 5.45: Histograma representando a distribuição *a posteriori* do número de intervalos para o modelo $\mathcal{M}_1$.

Na Tabela 5.22 são apresentadas as estimativas *a posteriori* para a fração de curados por tipo de tratamento pelos modelos $\mathcal{M}_0$ e $\mathcal{M}_1$. Para o modelo $\mathcal{M}_1$, a fração de cura para pacientes hemofílicos tratados com alguma dose é 22,35% e 87,65% em pacientes sem nenhuma dose. Podemos dizer que pacientes sem nenhuma dose de produtos derivados do sangue tem uma probabilidade de aproximadamente 88% de não contrair HIV-1. Observamos também,

que as estimativas associadas ao modelo $\mathcal{M}_0$ são bastante similares aquelas fornecidas pelo modelo $\mathcal{M}_1$.

Tabela 5.22: Resultados descritivos da fração de cura por tipo de tratamento, para os modelos $\mathcal{M}_0$ e $\mathcal{M}_1$, com $m^{'} = 40$.

Modelo	Tratamento	Média	Desvio padrão	HPD 95%
$\mathcal{M}_0$	Com dose	0,8771	0,0216	[0,8341;0,9183]
	Sem dose	0,2224	0,0230	[0,1757;0,2666]
$\mathcal{M}_1$	Com dose	0,8765	0,0214	[0,8317;0,9155]
	Sem dose	0,2234	0,0235	[0,1772;0,2690]

As funções de sobrevivência populacionais estimadas para os modelos $\mathcal{M}_0$ e $\mathcal{M}_1$, respectivamente, são apresentadas nas Figuras 5.46 e 5.47. Como comparação, apresentamos também a estimativa da função de sobrevivência populacional via estimador de Turnbull. Observamos que os resultados foram bastantes similares, sugerindo que os modelos estão bem ajustados aos dados. No entanto, o modelo $\mathcal{M}_1$ se aproxima um pouco mais da curva do estimador de Turnbull, quando avaliamos o grupo com alguma dose. Observamos também que, em ambas as figuras, o platô estabiliza nos valores da fração de cura apresentados na Tabela 5.22, aproximadamente.

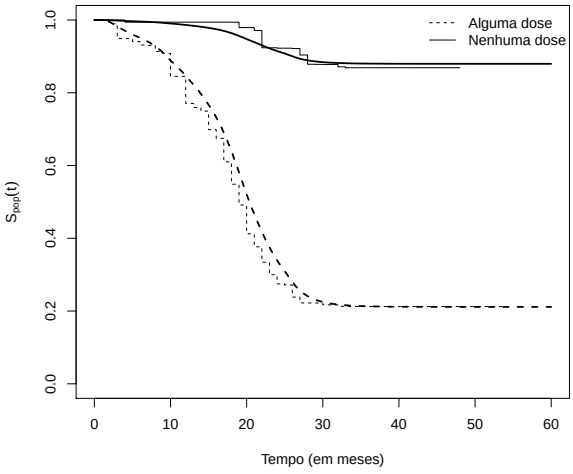

Figura 5.46: Estimativa da função de sobrevivência populacional para o modelo $\mathcal{M}_0$ (linha mais suave) e estimador de Turnbull (linha menos suave).

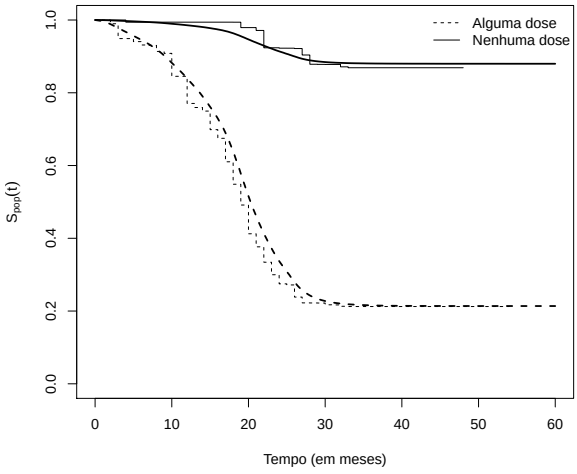

Figura 5.47: Estimativa da função de sobrevivência populacional para o modelo $\mathcal{M}_1$ (linha mais suave) e estimador de Turnbull (linha menos suave).

Capítulo 6

Considerações finais e trabalhos futuros

Apresentamos nesta tese duas propostas de extensões do MEP com grade aleatória para o ajuste de dados de sobrevivência sujeitos a censura intervalar envolvendo dados sem e com fração de cura. Estendemos o modelo dinâmico proposto por Demarqui et al. (2012) para dados sem fração de cura e o modelo introduzido por Demarqui et al. (2014) para o ajuste de dados de sobrevivência com fração de cura. Foi realizada uma análise de sensibilidade para avaliar o efeito do número máximo de intervalos admitidos *a priori*, bem como a especificação da grade inicial, sobre a qualidade de ajuste dos modelos.

Para os modelos sem fração de curados, vimos que as versões dinâmicas propostas por Gamerman (1991) e Demarqui et al. (2012) são casos particulares da modelagem proposta nesta tese quando condicionamos nos pseudotempos de falha. Ao ajustar os modelos propostos, aos dados de câncer de mama (Finkelstein, 1986), observamos que a extensão do modelo de Gamerman (1991) para dados sujeitos a censura intervalar, obteve o melhor ajuste aos dados, quando comparado aos demais modelos. Podemos perceber também que a abordagem dinâmica apresenta melhores resultados quando comparada aos modelos com efeito fixo no tempo. Entre os modelos propostos, vimos que $m^{'} = 20$ é o tamanho de grade admitido *a priori* preferível entre os modelos com grade inicial especificada pela grade do Tipo 1. Por outro lado, entre os modelos disponíveis na literatura, $m^{'} = 41$ é o tamanho de grade que mais se adequou aos dados.

De maneira geral, vimos que, ao introduzir uma estrutura de agrupamento para modelar a grade dos tempos do MEP, obtivemos melhores resultados em relação ao modelo com grade fixa. Sobre a estrutura dinâmica, vimos que o fator de desconto se adapta a medida que o número máximo de intervalos admitidos *a priori* varia, sempre compensando quando há a presença de mais ou menos dados a cada agrupamento de intervalo. Os ajustes para os dados de soro conversão de HIV-1 em pacientes hemofílicos, indicaram o modelo com fração de cura dinâmico com grade aleatória ($m^{'} = 40$) e grade especificada nos valores observados, é o que melhor se ajusta aos dados quando comparado ao caso com grade fixa. Neste modelo, os dados indicam que o número de intervalos mais provável é 22, com 40 intervalos admitidos *a priori*. Observamos também, que para o grupo tratado com "nenhuma dose" a fração de

cura estimada foi de 87,65%, indicando que aproximadamente 88% da população de pacientes hemofílicos tratados com nenhuma dose nunca irão desenvolver o HIV-1, conforme também indicado na estimativa da função de sobrevivência populacional.

Mostramos para os dados simulados sem fração de cura que os modelos dinâmicos propostos $\mathcal{M}_4$ e $\mathcal{M}_5$ conseguem retornar de maneira satisfatória o efeito real da covariável que apresenta um comportamento que varia no tempo além de apresentam um comportamento bem similar para com a função de sobrevivência real.

Com respeito aos dados com fração de curados gerados artificialmente, observamos, assim como na aos dados reais com fração de cura, que introduzir a estrutura de agrupamento do MPP para modelar a grade dos tempos na modelagem dinâmica nos permite obter o modelos dinâmicos com grade aleatória podemos conseguir uma melhor qualidade de ajuste aos dados. Para esses ajustes precisamos ressaltar que foi preciso reduzir a variabilidade na distribuição *a priori* para ψ_0, para melhorar a convergência das cadeias e garantir uma estabilidade nas estimativas dos parâmetros de interesse.

Um outro ponto a ressaltar é a respeito do limiar de tempo que define a origem do platô na curva de sobrevivência estimada pelo método de Tunrbull. Avaliamos somente um valor como limiar. Futuramente, iremos realizar uma análise de sensibilidade para este limiar poderia apresentar outros resultados para os critérios de seleção de modelos, favorecendo ou não o modelo dinâmico com grade aleatória.

Em linhas gerais, percebemos que com a utilização do algoritmo de ampliação de dados, podemos propor sempre novos modelos para dados de sobrevivência sujeitos a censura à direita e aplica-los para situações de censura intervalar.

Existem muitas propostas para trabalhos futuros envolvendo os modelos desenvolvidos. Uma delas, consiste em estender o modelo dinâmico com grade aleatória, sem covariáveis, proposto por Demarqui (2010), para fator de desconto aleatório e dados de sobrevivência sem fração de cura e sujeitos a censura intervalar.

No cenário de fração de curados com dados sujeitos a censura intervalar podemos obter algumas extensões. Por exemplo:

- Estender o modelo de mistura padrão (Berkson & Gage, 1952) baseando-se no modelo dinâmico descrito na Seção 4.2.

- Utilizar outras distribuições de probabilidade para $\mathbf{M}$, como por exemplo, a distribuição Binomial negativa (de Castro et al., 2009) e COM-Poisson (Rodrigues et al., 2009b), entre outras.

- Na disponibilidade de informações espaciais, podemos propor uma extensão ao modelo de Banerjee & Carlin (2004), incluindo uma fragilidade espacial no modelo dinâmico com fração de cura e grade aleatória para dados sujeitos a censura intervalar.

Nos modelos com fração de cura, vimos que ao introduzir a estrutura dinâmica para correlacionar as taxas do MEP, assim como da estrutura de agrupamento do MPP para modelar

a grade τ, a qualidade de ajuste melhora. Dessa forma, podemos realizar uma comparação do modelo de fração de cura dinâmico com grade aleatória proposto, com os modelos de Kim et al. (2006) e Demarqui et al. (2014), estendendo-os para dados de sobrevivência sujeitos a censura intervalar.

Apêndice A

Resultados das aplicações

A.1 Dados reais

Apresentamos a seguir, os resultados da análise de dados reais, dados de pacientes com câncer de mama, referente aos modelos $\mathcal{M}_4$ e $\mathcal{M}_5$.

Tabela A.1: Valores dos critérios LPML para os modelos $\mathcal{M}_4$ e $\mathcal{M}_5$ para grade tipo 2.

Fator de desconto	LPML							
	τ fixo	τ aleatório	τ fixo	τ aleatório	τ fixo	τ aleatório	τ fixo	τ aleatório
	$m^{'} = 10$		$m^{'} = 20$		$m^{'} = 30$		$m^{'} = 41$	
	Grade tipo 2							
0,1	299.03	300.49	304.57	301.92	310.43	302.91	313.73	306.44
0,15	297.19	298.26	300.25	298.75	308.58	301.80	311.95	302.78
0,2	294.99	297.92	298.98	297.28	302.97	299.12	308.93	300.47
0,25	293.83	296.66	296.55	296.05	300.20	296.76	305.93	298.51
0,3	292.47	296.32	295.51	294.84	298.27	295.70	303.91	297.02
0,35	291.66	296.00	293.84	294.17	296.70	294.70	300.88	295.45
0,4	291.20	296.06	292.84	294.08	295.55	294.10	297.82	294.39
0,45	290.92	296.21	291.65	293.82	293.49	293.81	295.82	293.65
0,5	290.96	296.14	291.56	293.68	292.10	292.68	293.68	292.58
0,55	291.18	296.54	291.24	294.08	291.32	292.67	292.54	291.93
0,6	291.69	296.67	291.49	294.78	290.60	292.97	291.15	291.45
0,65	292.32	297.36	291.99	295.11	290.25	293.08	290.12	291.31
0,7	293.21	297.87	293.00	295.79	289.97	293.62	289.32	291.92
0,75	294.11	298.46	294.25	297.89	290.47	294.69	288.73	292.45
0,8	295.48	299.13	296.52	300.38	291.45	296.50	288.89	293.92
0,85	296.79	299.85	299.07	301.59	293.33	298.06	290.18	295.98
0,9	298.31	300.84	301.80	302.28	296.21	300.56	293.06	298.45
0,95	300.16	301.78	304.92	305.06	301.11	303.67	297.71	302.29

Tabela A.2: Valores dos critérios WAIC para os modelos $\mathcal{M}_4$ e $\mathcal{M}_5$ para grade tipo 2.

Fator	WAIC							
de	τ fixo	τ aleatório	τ fixo	τ aleatório	τ fixo	τ aleatório	τ fixo	τ aleatório
desconto	$m' = 10$		$m' = 20$		$m' = 30$		$m' = 41$	
	Grade tipo 2							
0,1	297,87	299,26	299,76	299,21	303,65	304,51	307,12	302,68
0,15	295,97	297,81	298,03	297,49	302,47	304,51	306,46	300,61
0,2	294,18	296,98	297,07	296,30	300,94	304,51	305,42	298,91
0,25	293,17	296,36	295,56	295,50	298,88	304,51	303,23	297,44
0,3	292,12	296,08	294,52	294,41	297,30	295,14	301,40	296,12
0,35	291,47	295,84	293,35	293,87	295,73	294,19	298,78	294,72
0,4	291,06	295,86	292,55	293,81	294,38	293,75	296,92	293,84
0,45	290,81	295,96	291,56	293,58	293,20	293,30	295,07	292,84
0,5	290,86	296,08	291,47	293,53	291,97	293,30	293,31	292,14
0,55	291,13	296,50	291,17	293,89	291,27	293,30	292,30	291,72
0,6	291,64	296,63	291,39	294,46	290,54	293,30	290,84	291,22
0,65	292,30	297,33	291,88	294,89	290,15	293,30	289,98	291,20
0,7	293,20	297,86	292,89	295,69	289,86	293,30	289,19	291,73
0,75	294,10	298,46	294,15	297,52	290,34	293,30	288,58	292,31
0,8	295,47	299,12	296,30	299,48	291,22	293,30	288,73	293,76
0,85	296,79	299,85	298,72	300,13	293,04	293,30	290,01	295,72
0,9	298,31	300,84	301,39	302,01	295,96	293,30	292,55	298,24
0,95	300,16	301,78	304,40	304,51	300,74	293,30	297,55	302,01

Tabela A.3: Valores dos critérios DIC para os modelos $\mathcal{M}_4$ e $\mathcal{M}_5$ para grade tipo 2.

Fator	DIC							
de	τ fixo	τ aleatório	τ fixo	τ aleatório	τ fixo	τ aleatório	τ fixo	τ aleatório
desconto	$m' = 10$		$m' = 20$		$m' = 30$		$m' = 41$	
	Grade tipo 2							
0,1	303.17	312.48	307.49	307.19	314.03	308.09	316.06	309.76
0,15	301.33	310.82	305.39	305.60	310.81	305.95	315.13	307.36
0,2	298.55	309.81	303.32	304.31	308.40	304.31	314.30	305.21
0,25	297.62	309.22	301.88	303.43	305.34	302.02	311.50	303.42
0,3	296.56	309.10	300.29	301.95	303.96	300.81	309.11	302.06
0,35	295.63	308.33	298.60	302.08	301.67	299.89	305.76	300.42
0,4	294.87	308.73	297.74	301.37	300.04	300.07	302.91	298.62
0,45	294.28	307.82	297.16	303.76	299.09	300.21	301.49	297.88
0,5	294.30	307.29	297.15	303.42	297.36	298.11	298.58	298.44
0,55	294.28	306.74	298.05	303.99	296.67	300.57	297.36	297.22
0,6	294.46	305.58	300.77	311.47	295.94	303.98	296.16	296.63
0,65	295.04	305.56	304.43	310.94	297.27	303.75	294.59	297.15
0,7	295.80	304.76	310.55	309.30	297.59	304.78	294.53	301.44
0,75	296.33	303.95	315.40	323.86	300.90	308.64	293.56	302.32
0,8	297.63	303.45	329.20	338.39	308.05	318.76	295.25	307.22
0,85	298.59	303.22	339.66	323.50	314.45	318.42	300.90	316.31
0,9	299.90	303.25	344.47	329.36	318.36	324.15	308.51	318.36
0,95	301.46	303.61	351.46	341.44	332.92	333.21	316.76	330.33

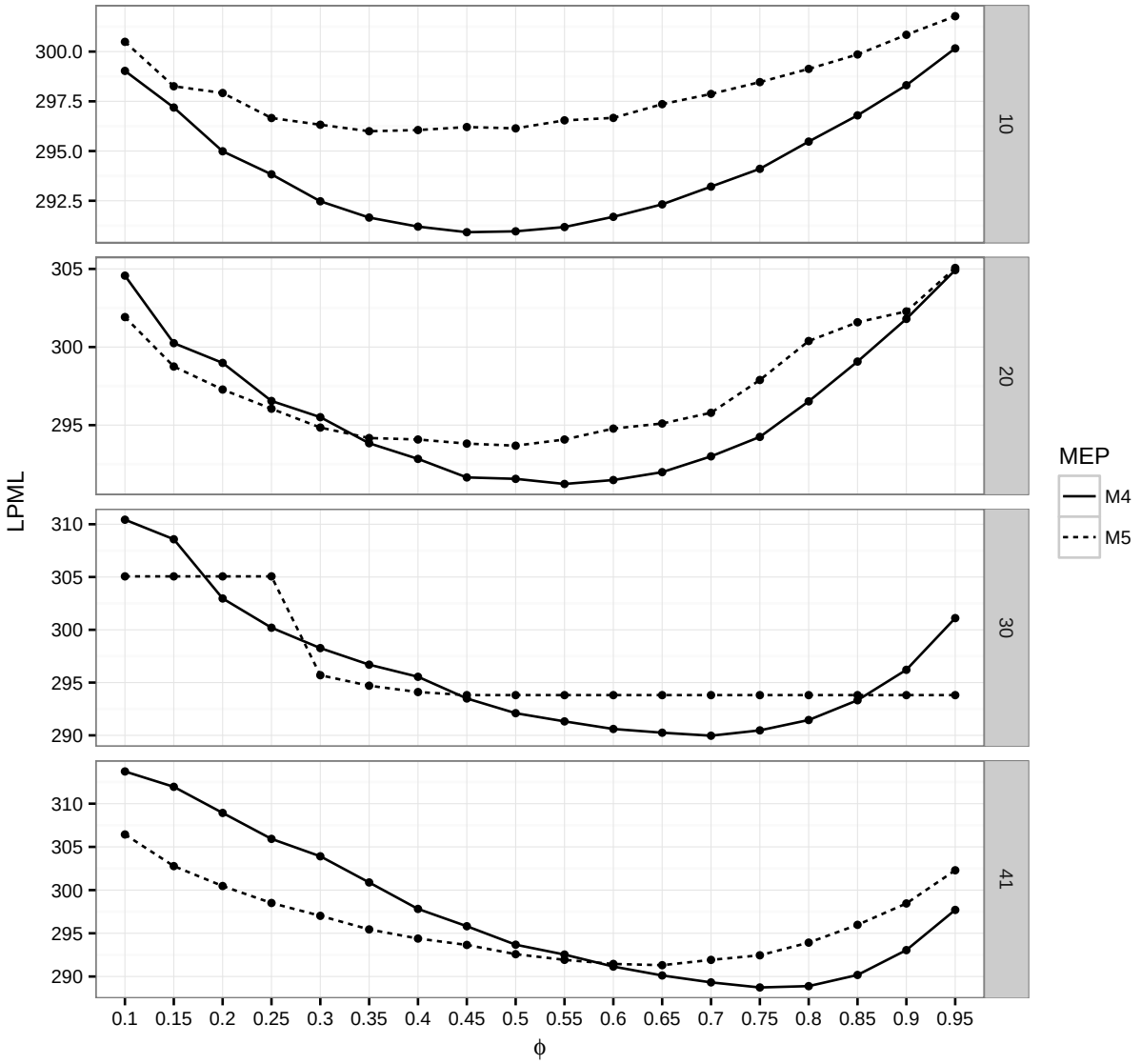

Figura A.1: Comparação dos modelos $\mathcal{M}_4$ e $\mathcal{M}_5$, via LPML, para os diferentes valores ϕ, $m^{'}s$ e grade tipo 2.

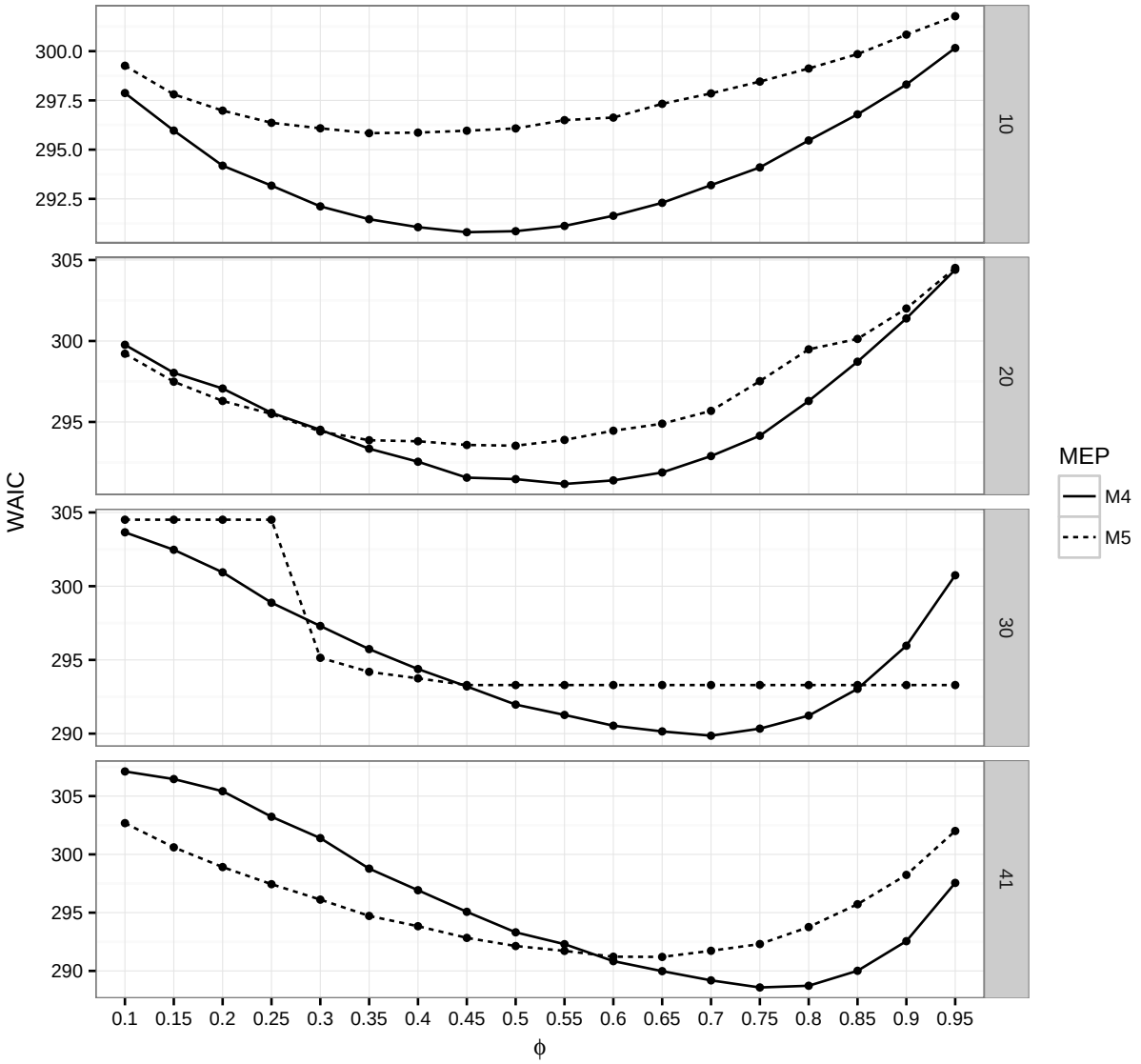

Figura A.2: Comparação dos modelos $\mathcal{M}_4$ e $\mathcal{M}_5$, via WAIC, para os diferentes valores ϕ, $m^{'}s$ e grade tipo 2.

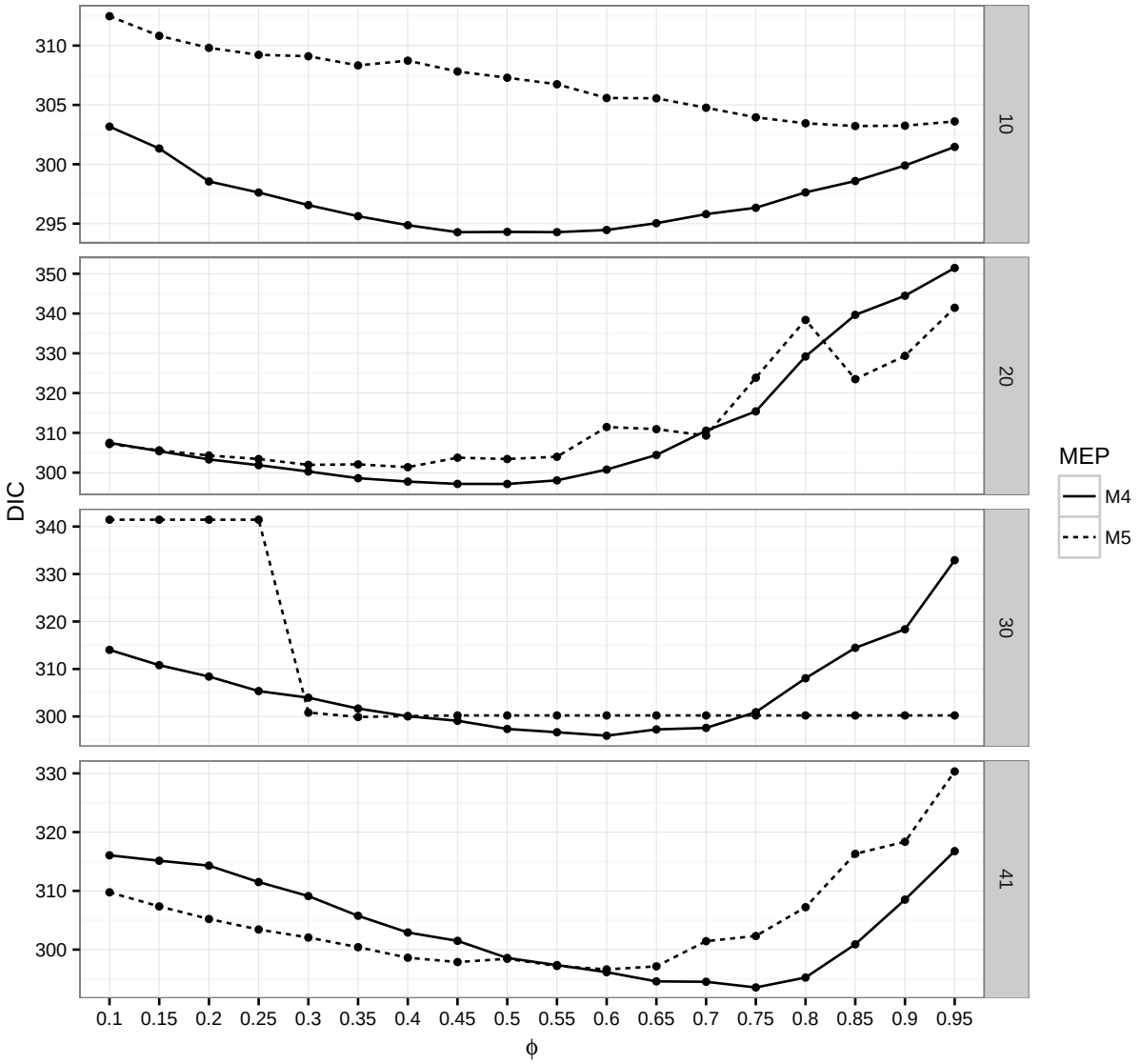

Figura A.3: Comparação dos modelos $\mathcal{M}_4$ e $\mathcal{M}_5$, via LPML, para os diferentes valores ϕ, $m^{'}s$ e grade tipo 2.

Apêndice B

Códigos - Geração dos dados artificiais

Neste Apêndice serão apresentados os códigos utilizados para a obtenção dos dados artificiais.

B.1 Dados sem fração de cura

```
#-----------------------------------------------------
rm(list=ls(all=TRUE))

set.seed(1234567890)

#--- Função com o efeito variando no tempo:
time <- function(t, alpha, gama, cov1, unif.val){
if(t <= 1.5){
elimpred <- exp(0.6*cov2)
f <- -alpha*(t^gama)*elimpred-log(unif.val)
return(f)
}else{
elimpred <- exp(1.4*cov2)
f <- -alpha*(t^gama)*elimpred-log(unif.val)
return(f)
}
}

#--- Especificações da amostra:

n <- 500
alpha <- 0.3
gama  <- 1.5
```

```
u <- runif(n)
x1 <- rbinom(n,1,0.5)

t.falha <- rep(NA, n)

for(i in 1:n){
raiz <- uniroot(time, c(0, 10000), alpha=alpha,
 gama=gama, cov1=x1[i], unif.val=u[i])
t.falha[i] <- raiz$root
}

#--- Tempos observados:

t.cens <- rexp(n, 0.1)
tempo  <- pmin(t.falha, t.cens)
delta  <- ifelse(t.falha<=t.cens, 1, 0)

#--- Obtendo os intervalos

L <- rep(NA, length(tempo))
R <- rep(NA, length(tempo))

for(i in 1:n){
if(delta[i]==0){
L[i] <- tempo[i]
R[i] <- Inf
}else{
L[i] <- 0
soma <- runif(1, 0.1, 0.5)
R[i] <- soma
while((L[i] <= tempo[i] & tempo[i] < R[i])==FALSE){
L[i] <- L[i] + soma
soma <- runif(1, 0.1, 0.5)
R[i] <- R[i] + soma
}
}
}

dados <- as.data.frame(cbind(L, tempo, R, delta, x1))
names(mydata) <- c("left","time","right", "delta", "x1")
```

```
#---------------------------------------------------
```

B.2 Dados com fração de cura

```
#---------------------------------------------------
rm(list=ls(all=TRUE))

set.seed(1234567890)

#--- Especificações da amostra:

n <- 500
lambda <- 0.03
shape <- 1.5

beta0     <- 0.45
beta1     <-  -2.5
beta      <- c(beta0,beta1)
beta.real <- beta

x1 <- rbinom(n,1,0.5)
X <- cbind(1,x1)

#--- Preditor linear:

theta <- exp(X%*%beta)
mi    <- rpois(n, lambda = theta)

time <- rep(0, n)

#--- Gerando os tempos de promoção via distribuição de Weibull:
for( i in 1:n ){
if( mi[i]==0 ){
time[i] <- Inf # Cured indivíduals
}else{
pro.time <- c()
for(m in 1:mi[i]){
u <- runif(1)
```

```
aux <- (-log(u)/lambda)^(1/shape)
pro.time <- c(pro.time, aux)
}
time[i] <- min(pro.time)
}
}

#--- tempos de censura:
a <- 27
b <- 15
C <- pmin(a, b*rexp(n, rate=0.1))

#--- Tempos observados:

t     <- pmin(time, C)
delta <- ifelse(time <= C, 1, 0)

#--- Obtendo os intervalos

L <- rep(NA, length(tempo))
R <- rep(NA, length(tempo))

for(i in 1:n){
if(delta[i]==0){
L[i] <- tempo[i]
R[i] <- Inf
}else{
L[i] <- 0
soma <- runif(1, 0.1, 0.5)
R[i] <- soma
while((L[i] <= tempo[i] & tempo[i] < R[i])==FALSE){
L[i] <- L[i] + soma
soma <- runif(1, 0.1, 0.5)
R[i] <- R[i] + soma
}
}
}

dados <- as.data.frame(cbind(L, tempo, R, delta, x1))
names(mydata) <- c("left","time","right", "delta", "x1")
```

```
#-------------------------------------------------------
```

Referências Bibliográficas

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

Banerjee, M., & Carlin, B. P. (2004). Parametric spatial cure rate models for interval-censored time-to-relapse data. *Biometrics*, *60*, 268–275.

Barry, D., & Hartigan, J. A. (1992). Product partition models for change point problems. *Annals of Statistics*, *20*, 260–279.

Barry, D., & Hartigan, J. A. (1993). A Bayesian analysis for change point problems. *Journal of the American Statistical Association*, *88*, 309–319.

Berkson, J., & Gage, R. P. (1952). Survival curvese for cancer patients following treatment. *Journal of the American Statistical Association*, *47*, 501–515.

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistic Society B*, *11*, 15–53.

Breslow, N. E. (1974). Covariance analysis of censored survival data. *Biometrics*, *30*, 89–99.

Chen, D., Sun, J., & Peace, K. E. (2013). *Interval-censored Time-to-event Data: Methods and Applications*. Boca Raton, FL: CRC Press.

Chen, M. H., & Ibrahim, J. G. (2001). Maximum likelihood methods for cure rate models with missing covariates. *Biometrics*, *57*, 43–52.

Chen, M.-H., Ibrahim, J. G., & Sinha, D. (1999). A new Bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, *94*, 909–919.

Collett, D. (2015). *Modelling Survival Data in Medical Research, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press.

Colosimo, E., & Giolo, S. (2006). *Análise de Sobrevivência Aplicada*. São Paulo: Edgard Blücher.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society*, *34*, 187–188.

da Costa, J. C. B. (2016). *Modelos semiparamétricos de fração de cura para dados com censura intervalar*. Dissertação de mestrado, Universidade de São Paulo - IME - USP.

de Castro, M., Cancho, V., & Rodrigues, J. (2009). A Bayesian long-term survival model parametrized in the cured fraction. *Biometrical Journal*, *51*, 443–455.

Dellaportas, P., & Smith, A. F. M. (1993). Bayesian inference for generalized linear and proportional hazards models via gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *42*, 443–459.

Demarqui, F. N. (2008). *Modelo Exponencial por partes via Model Partição Produto*. Dissertação de mestrado, Universidade Federal de Minas Gerais.

Demarqui, F. N. (2010). *Uma classe mais flexível de modelos semiparamétricos para dados de sobrevivência*. Tese de doutorado, UFMG.

Demarqui, F. N., Dey, D. K., Loschi, R. H., & Colosimo, E. A. (2014). Fully semiparametric Bayesian approach for modeling survival data with cure fraction. *Biometrical Journal*, *56*, 198–218.

Demarqui, F. N., Loschi, R. H., & Colosimo, E. A. (2008). Estimating the grid of time-points for the piecewise exponential model. *Lifetime Data Analysis*, *14*, 333–356.

Demarqui, F. N., Loschi, R. H., Dey, D. K., & Colosimo, E. A. (2012). A class of dynamic piecewise exponential models with random time grid. *Journal of Statistical Planning and Inference*, *142*, 728–742.

Dey, D. K., Chen, M.-H., & Chang, H. (1997). Bayesian approach for nonlinear random effects models. *Biometrics*, *53*, 1239–1252.

Dorey, F. J., Little, R. J., & Schenker, N. (1993). Multiple imputation for thereshold-crossing data with interval censoring. *Statistics in Medicine*, *12*, 1589–1603.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, *38*, 1041–1046.

Farewell, V. T. (1986). Mixture models in survival analysis: Are they worth the risk? *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, *14*, 257–262.

Fay, M. P., & Shaw, P. A. (2010). Exact and asymptotic weighted logrank tests for interval censored data: The interval R package. *Journal of Statistical Software*, *36*, 1–34.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, *42*, 845–854.

Finkelstein, D. M., & Wolfe, R. A. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics*, *41*, 933–945.

Friedman, M. (1982). Piecewise exponential models for survival data with covariates. *The Annals of Statistics*, *10*, 101–113.

Gamerman, D. (1991). Dynamic Bayesian models for survival data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *40*, 63–79.

Gamerman, D. (1994). Bayes estimation of the piece-wise expoential distribution. *IEEE Transactions on Reliability*, *43*, 128–131.

Gelfand, A. E., Dey, D. K., & Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. *Bayesian Statistics*, *4*, 147–167.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398–409.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian Data Analysis, Second Edition (Chapman & Hall/CRC Texts in Statistical Science)*. Boca Raton, FL: Chapman and Hall/CRC, 2 ed.

Gelman, A., Hwang, J., & Vehtari, A. (2013). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*, 997–1016.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Geweke, J. (1992). [statistics, probability and chaos]: Comment: Inference and prediction in the presence of uncertainty and determinism. *Statistical Science*, *7*, 94–101.

Gilks, W., & Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society: Series C*, *41*, 337–348.

Gilks, W. R., Best, N. G., & Tan, K. K. C. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Applied Statistics*, *4*, 455–472.

Gómez, G., M., C., & Oller, R. (2004). Frequentist and Bayesian approaches for interval-censored data. *Statistical Papers*, *45*, 139–173.

Goedert, J. J., Kessler, C. M., Aledort, L. M., Biggar, R. J., Andes, W. A., White, G. C. I., Drummond, J. E., Vaidya, K., Mann, D. L., Eyster, M. E., Ragni, M. V., Lederman, M. M., Cohen, A. R., Bray, G. L., Rosenberg, P. S., Friedman, R. M., Hilgartner, M. W., Blattner, W. A., Kroner, B., & Gail, M. H. (1989). A prospective study of human immunodeficiency virus type 1 infection and the development of aids in subjects with hemophilia. *New England Journal of Medicine*, *321*, 1141–1148.

Goetghebeur, E., & Ryan, L. (2000). Semiparametric regression analysis of interval-censored data. *Biometrics*, *56*, 1139–1144.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*, 711–732.

Hashimoto, E. M., Ortega, E. M. M., Cordeiro, G. M., & Cancho, V. G. (2015). A new long-term survival model with interval-censored data. *Sankhya B*, *77*, 207–239.

Heidelberger, P., & Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, *31*, 1109–1144.

Henschel, V., Engel, J., Hölzel, D., & Mansmann, U. (2009). A semiparametric Bayesian proportional hazards model for interval censored data with frailty effects. *Medical Research Methodology*, *9*.

Hoel, D. G., & Walburg, H. E. (1972). Statistical analysis of survival experiments. *Journal of the National Cancer Institute*, *49*, 361–372.

Hu, T., & Xiang, L. (2013). Efficient estimation for semiparametric cure models with interval-censored data. *Journal of Multivariate Analysis*, *121*, 139 – 151.

Huang, J. (1996). Efficient estimation for the Cox model with interval censoring. *The Annals of Statistics*, *24*, 540–568.

Ibrahim, J. G., Chen, M. H., & Sinha, D. (2001). *Bayesian Survival Analysis*. New York, NY: Springer.

Kalbfleisch, J. D. (1978). Non-parametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society. Series B (Methodological)*, *40*(2), 214–221.

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley.

Kalbfleisch, R. L., J. D. e Prentice (1973). Marginal likelihoods based on cox´s regression and life models. *Biometrika*, *60*, 267–278.

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

Kim, J. S. (2003). Efficient estimation for the proportional hazards model with left-truncated and case 1 interval-censored data. *Statistica Sinica*, *13*, 519–537.

Kim, J. S., & Proschan, F. (1991). Piecewise exponential estimator of the survival function. *IEEE Transactions on Reliability*, *40*, 134–139.

Kim, M. Y., Gruttola, V. G. D., & Lagakos, S. W. (1993). Analyzing doubly censored data with covariates, with application to AIDS. *Biometrics*, *49*, 13–22.

Kim, S., Chen, M. H., Dey, D., & Gamerman, D. (2006). Bayesian dynamic models for survival data with a cure fraction. *Lifetime Data Analysis*, *13*, 17–35.

Kim, Y.-J., & Jhun, M. (2008). Cure rate model with interval censored data. *Satatistics in Medicine*, *27*, 3–14.

Klein, J. P., & Moechsberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Verlag, New York.

Klein, J. P., Van Houwelingen, H. C., Ibrahim, J. G., & Scheike, T. H. (2013). *Handbook of Survival Analysis*. Boca Raton, FL: CRC Press.

Kongerud, J., & Samuelsen, S. (1991). A longitudinal study of respiratory symptoms in aluminum potroom workers. *American Review of Respiratory Diseases*, *144*, 10–16.

Kroner, B., Rosenberg, P., Adedort, L., Alvord, W., & Goedert, J. (1994). HIV-1 infection incidence among people with hemophilia in the United States and Western Europe, 1978-1990. *Journal of Acquired Immune Deficiency Syndromes*, *7*, 279–286.

Kuk, A. Y. C., & Chen, C.-H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, *79*, 531–541.

Lam, K. F., Wong, K. Y., & Zhou, F. (2013). A semiparametric cure model for interval-censored data. *Biometrical Journal*, *55*, 771–788.

Lindsey, J. C., & Ryan, L. M. (1998). Tutorial in biostatistics: Methods for interval-censored data. *Statistics in Medicine*, *17*, 219–238.

Liu, H., & Shen, Y. (2009). A semiparametric regression cure model for interval-censored data. *Journal of the American Statistical Association*, *104*, 1168–1178.

Liu, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, *89*, 958–966.

Loschi, R. H., & Cruz, F. R. B. (2005). Extension to the product partition model: computing the probability of a change. *Computational Statistics & Data Analysis*, *24*, 305–319.

Maller, X., R. A. & Zhou (1996). *Survival Analysis with Long-Term Survivors*. New York, NY: Wiley.

Mckeague, I. W., & Tighiouart, M. (2000). Bayesian estimators for conditional hazard functions. *Biometrics*, *56*, 1007–1015.

McMahan, C. S., & Wang, L. (2014). *ICsurv: A package for semiparametric regression analysis of interval-censored data*. R package version 1.0.

Odell, P. M., Anderson, K. M., & D'Agostinho, R. B. (1992). Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics*, *48*, 951–959.

Pan, W. (2000). A multiple imputation approach to Cox regression with interval censored data. *Biometrics*, *56*, 199–203.

Peto, R. (1973). Experimental survival curves for interval-censored data. *Applied Statistics*, *22*, 86–91.

Petris, G. (2010). An R Package for Dynamic Linear Models. *Journal of Statistical Software*, *36*(12), 1 – 12.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, *6*, 7–11.

Quintana, F. A., & Iglesias, P. L. (2003). Nonparametric Bayesian clustering and product partition models. *Journal of the Royal Statistics Society B*, *2*, 557–574.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rücker, G., & Messerer, D. (1988). Remission duration: an example of interval censored observation. *Statistics in Medicine*, *7*, 1139–1145.

Rodrigues, J., Cancho, V. G., de Castro, M., & Louzada-Neto, F. (2009a). On the unification of long-term survival models. *Statistics & Probability Letters*, *79*, 753 – 759.

Rodrigues, J., de Castro, M., Cancho, V. G., & Balakrishnan, N. (2009b). COM-Poisson cure rate survival models and an application to a cutaneous melanoma data. *Journal of Statistical Planning and Inference*, *139*, 3605 – 3611.

Rodriguez, P. P. (2014). *ars: Adaptive Rejection Sampling - Original C++ code from Arnost Komarek based on ars.f written by P. Wild and W. R. Gilks*. R package version 0.5. URL https://CRAN.R-project.org/package=ars

Sahu, S. K., Dey, D. K., Aslanidou, D. K., & Sinha, D. (1997). A Weibull regression model with gamma frailties for multivariate survival data. *Lifetime Data Analysis*, *3*, 123–137.

Samuelsen, S. O., & Kongerud, J. (1993). Evaluation of applying interval censoring on longitudinal data on asthmatic symptoms. Tech. rep., Statistical Research Report 2, Institute of Mathematics, University of Oslo.

Seaman, S. R., & Bird, S. M. (2001). Proportional hazards model for interval-censored failure times and time-dependent covariates: application to hazard of hiv infection of injecting drug users in prison. *Statistics in Medicine*, *20*, 1855–1870.

Sinha, D., Chen, M.-H., & Ghosh, S. K. (1999). Bayesian analysis and model selection for interval-censored survival data. *Biometrics*, *55*, 585–590.

Smith, P. J., Thompson, T. J., & Jereb, J. A. (1997). A model for interval censored tuberculosis outbreak data. *Statistics in Medicine*, *16*, 485–496.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*, 583–639.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*, 485–493.

Sun, J. (1996). A nonparametric test for interval-censored failure time data with application to AIDS studies. *Statistics in Medicine*, *15*, 1387–1395.

Sun, J. (1997). Regression analysis of interval-censored failure time data. *Statistics in Medicine*, *16*, 497–504.

Sun, J. (2006). *The Statistical Analysis of Failure Time Data*. New York, NY: Wiley.

Sy, J., & Taylor, J. (2001). Standard errors for the Cox proportional hazards cure model. *Mathematical and Computer Modelling*, *33*, 1237–1251.

Tanner, M. A. (1991). *Tools for Statistical Inference: Observed Data and Data Augmentation Methods*. New York, NY: Springer-Verlag.

Tanner, M. A., & Wong, W. H. (1987). The application of imputation to an estimation problem in grouped lifetime analysis. *Technometrics*, *29*, 23–32.

Thompson, L. A., & Chhikara, R. S. (2003). A Bayesian cure rate model for repeated measurements and interval censoring. Proceedings of JSM.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, *38*, 290–295.

van Dyk, D. A., & Meng, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics*, *10*, 1–50.

Wang, X., Chen, M.-H., & Yan, J. (2013). Bayesian dynamic regression models for interval censored survival data with application to children dental health. *Lifetime Data Analysis*, *19*, 297–316.

Wang, X., Yan, J., & Chen, M.-H. (2014). *dynsurv: Dynamic models for survival data*. R package version 0.2-2.

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.

Wei, G. C. G., & Tanner, M. A. (1991). Application of multiple imputation to the analysis of censored regression data. *Biometrics*, *47*, 1297–1309.

West, M., & Harrison, P. J. (1997). *Bayesian Forecasting & Dynamic Models*. New York: Springer, 2nd ed.

Yakovlev, A., Asselain, B., Bardou, V., A.Fourquet, Houag, T., Rochefediere, A., & Tsodikov, A. (1993). A simple stochastic model of tumor recurrence and its application to data on premenopausal breast cancer. *Biometrie et Analyse de Donnees Spatio-Temporalles*, *12*, 67–82.

Zhang, Z., & Sun, J. (2010). Interval censoring. *Statistical Methods in Medical Research*, *19* , 53–70.

Zhao, X., Lim, H., & Sun, J. (2005). Estimating equation approach for regression analysis of failure time data in the presence of interval-censoring. *Journal of Statistical Planning and Inference*, *129* , 145–157.