Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Modelos de Sobrevivência com fração de cura via Partição Bayesiana

Jhon F. Bernedo Gonzales

São Carlos
2014

Jhon F. Bernedo Gonzales

Modelos de Sobrevivência com fração de cura via Partição Bayesiana

Tese apresentada ao Departamento de Estatística da Universidade Federal de São Carlos - DEs/UFSCar como parte dos requisitos para obtenção do título de doutor em estatística.

Orientadores:

Prof. Dra. Vera Lucia Damasceno Tomazella

Prof. Dr. Mário de Castro Andrade Filho

São Carlos-SP

2014

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária/UFSCar

B525ms

Bernedo Gonzales, Jhon Franky.
Modelos de sobrevivência com fração de cura via partição bayesiana / Jhon Franky Bernedo Gonzales. -- São Carlos : UFSCar, 2014.
102 f.

Tese (Doutorado) -- Universidade Federal de São Carlos, 2014.

1. Análise de sobrevivência. 2. Modelos de partição. 3. Modelos de sobrevivência com fração de cura. 4. Série de potências. I. Título.

CDD: 519.9 (20ª)

UNIVERSIDADE FEDERAL DE SÃO CARLOS
Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Estatística
Via Washington Luís, Km 235 - C.P.676 - CGC 45358058/0001-40
FONE: (016) 3351-8292 – Email: ppgest@ufscar.br
13565-905 - SÃO CARLOS-SP - BRASIL

FOLHA DE APROVAÇÃO

Aluno(a) : Jhon Franky Bernedo Gonzales

TESE DE DOUTORADO DEFENDIDA E APROVADA EM 30/05/2014 PELA COMISSÃO JULGADORA:

Presidente _____
Profa. Dra. Vera Lucia Damasceno Tomazella (DEs-UFSCar / Orientadora)

1º Examinador _____
Prof. Dr. Mário de Castro Andrade Filho (ICMC-USP / Coorientador)

2º Examinador _____
Prof. Dr. Francisco Louzada Neto (ICMC-USP)

3º Examinador _____
Prof. Dr. Gustavo Leonel Gilardoni Avalle (UnB)

4º Examinador _____
Prof. Dr. Ricardo Sandes Ehlers (ICMC-USP)

5º Examinador _____
Profa. Dra. Rosangela Helena Loschi (UFMG)

Resumo

Em geral, os modelos para dados de sobrevivência com fracão de cura relacionam a fração de cura com as covariáveis por meio de diferentes funções de ligação, por exemplo, a função de ligação logito e não consideram o problema de seleção de covariáveis que tem um efeito na fração de cura. Assim neste trabalho é proposto uma modelagem que considera uma partição do espaço preditor em que a fração de cura depende localmente das covariáveis. Neste contexto, adota-se uma tesselação por hiperplanos ortogonais aos eixos a fim de obter uma partição do espaço preditor com a vantagem que os modelos propostos selecionam as covariáveis que têm efeito na fração de cura. A modelagem desenvolvida estende o modelo de partição bayesiana proposto por Hoggart & Griffin (2001) por incluir informações de variáveis qualitativas com mais de duas categorias e dessa forma uma nova estratégia computacional é considerada. Essa extensão permite capturar os efeitos das covariáveis numa estrutura local na qual considera-se que o número de causas competitivas segue distribuição série de potências. Esta distribuição é flexível pois inclui casos particulares, tais como a distribuição binomial, Poisson, binomial negativa e logarítmica. Para demonstrar o potencial da metodologia descrita, utilizou-se dois conjunto de dados relacionados com estudos de câncer.

Abstract

In general, models for survival data with a cure fraction relate the cure fraction with the covariates using different link functions, for example, the logit link function and do not consider the problem of selection of covariates that have an effect on the cure fraction. So, in this work we propose a model that considers a partition of the predictor space in which the cure fraction depends locally of covariates. In this context, it adopts a orthogonal hyperplane tessellation to the axes to obtain a partition of the predictor space with the advantage that the proposed model selects the covariates that have an effect on the cure fraction. The developed modeling extends the Bayesian partition model proposed by Hoggart & Griffin (2001) to include information for qualitative variables with more than two categories and therefore a new computational strategy is considered. This extension allows to capture the effects of covariates on a local structure in which it is considered that the number of competing causes follows a power series distribution. This distribution is flexible because it includes special cases such as the binomial, Poisson, negative binomial and logarithmic distributions. To demonstrate the potential of the methodology, we used two set of data relating with cancer studies.

Eu não procuro saber as respostas, procuro compreender as perguntas. Confúcio

Agradecimentos

Primeiramente agradeço a Deus, que me dá saúde e força para superar os obstáculos e provas todos os dias.

Ao meu pai, à minha mãe, pelo constante apoio e ânimo em minha vida, às minhas irmãs Hayme e Gleny pela compreensão e ajuda nos momentos difíceis.

À minha orientadora Vera Lucia D. Tomazella e ao meu coorientador Mário de Castro, pela orientação e incentivo na elaboração e condução do trabalho. Foi um prazer trabalhar com eles e são inspiração em minha vida para continuar estudando.

Aos professores do Departamento de Estatística da Universidade Federal de São Carlos, que me abriram as portas e me ofereceram ambiente acolhedor e sadio para que eu pudesse realizar meu doutorado.

Aos meus amigos Mauro e Paulo Henrique, por sua amizade, e a todos os meus amigos que sempre estiveram carinhosamente presentes, contribuindo com críticas, sugestões e paciente tolerância.

Finalmente, agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo auxílio concedido para este trabalho.

Sumário

Lista de Figuras

Lista de Tabelas

Tabela de Símbolos

p	Número de variáveis preditoras
$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$	variáveis preditoras
n	Tamanho da amostra
$\boldsymbol{\mathcal{X}}$	Espaço preditor
$\boldsymbol{\mathcal{T}}$	Tesselação por hiperplanos ortogonais
M	Número de regiões que produz a tesselação no espaço preditor
N	Número de riscos latentes de um indivíduo
$A_N(\cdot)$	Função geradora de probabilidade da variável aleatória N

Capítulo 1

Introdução

Em geral, estudos envolvendo observações até a ocorrência de um evento de interesse são numerosos e estão presentes em pesquisas da área médica, engenharia, de economia financeira, entre outras. Por exemplo, em estudos clínicos, o evento de interesse pode ser a recorrência do tumor, recidiva de uma doença, a morte do paciente, etc. Na área financeira, o evento de interesse pode ser o abandono de um cliente, o não pagamento de empréstimos, a ocorrência de um sinistro etc. Neste contexto, a análise de sobrevivência permite o estudo do tempo até a ocorrência do evento de interesse e, geralmente, este tempo é chamado tempo de falha. Porém, existem indivíduos da população em estudo em que não é observado o evento de interesse ao final do estudo.

Por exemplo, na área de estudos médicos existem alguns indivíduos que não apresentarão a recorrência de uma doença mesmo sendo acompanhados por um tempo suficientemente grande. De forma similar, na área financeira, existem uma proporção de clientes que não abandonam o banco num intervalo de um ano.

Dados observados em forma parcial ou incompleta são denominados censurados. Neste cenário, os dados de sobrevivência são em geral compostos por uma parte discreta que é definida pela variável indicadora de censura e uma parte contínua, que envolve o tempo de falha ou de censura. Assim, para a modelagem de dados de sobrevivência, o estimador de Kaplan-Meier (Kaplan & Meier, 1958) é usualmente empregado para estimar a função de sobrevivência de um ponto de vista não paramétrico.

O modelo proposto por Cox (1972), conhecido como modelo de riscos proporcionais, é geralmente aplicado a dados de sobrevivência considerando variáveis preditoras. As vantagens, desvantagens e extensões do modelo proposto por Cox (1972) são muito discutidos na literatura. Entre as referências que podem ser citadas estão Lawless (2002)

e Kalbfleisch & Prentice (2002).

Em estudos clínicos, particularmente em pesquisas de câncer de mama, pode ser visto em Farewell & Sprott (1986) e Peng & Dear (2000), câncer de cólon em Lambert *et al.* (2007) e melanoma em Chen *et al.* (1999), em geral assume-se que os pacientes são suscetíveis ao evento de interesse (por exemplo, morte ou recidiva da doença). Entretanto, na atualidade, com os avanços nos tratamentos de câncer e, por consequência na eficácia deles, os estudos conduzem a uma proporção de pacientes que não são suscetíveis ao evento de interesse esperado. Tem-se na literatura, técnicas estatísticas em análise de sobrevivência adequadas para essa situação, que mostram que esses indivíduos são "imunes" ao evento de interesse. A população da qual eles fazem parte possui uma fração de curados. Neste contexto, uma metodologia usada em análise de sobrevivência que considera uma proporção de curados é a de modelos de longa duração, também chamados de modelos com fração de cura.

Os modelos de análise de sobrevivência com longa duração possuem uma vantagem em relação aos modelos de sobrevivência usuais, no sentido de incorporarem a heterogeneidade de duas subpopulações (suscetíveis e imunes); são, por isso, conhecidos como modelos de mistura.

A suposição de que alguns pacientes nunca experimentarão o evento de interesse é baseada em considerações científicas ou empíricas, como a presença de um grande número de sobreviventes de longa duração (alta proporção de censura). O estimador Kaplan-Meier é uma boa forma de evidenciar essa presença de censura, uma vez que um grande número de censuras pode ser observado na cauda, ou seja, pode ser testada a existência de pacientes "curados". O gráfico desse estimador deve apresentar uma cauda em um nível aproximadamente constante e estritamente maior que zero, por um período de tempo considerável.

Neste capítulo, são apresentados: revisão bibliográfica, na Seção 1.1; objetivos, na Seção 1.2 e organização do trabalho na Seção 1.3.

1.1 Revisão bibliográfica

Existe uma extensa literatura sobre os modelos de longa duração em que os autores vêm discutindo a questão de modelos envolvendo misturas de distribuições. Dentre esses modelos, o trabalho pioneiro foi apresentado por Boag (1949), que utilizou o método de

máxima verossimilhança para estimar a proporção de sobreviventes em uma população de 121 mulheres com câncer de mama, esse experimento teve a duração de 14 anos. Baseado na ideia de Boag (1949), Berkson & Gage (1952) propuseram um modelo de mistura, com o objetivo de estimar a proporção de curados numa população submetida a um tratamento de câncer de estômago.

Farewell (1977) abordou o modelo de mistura Weibull e investigou como os fatores de risco (por exemplo idade ao primeiro parto) afetam o tempo de desenvolvimento do câncer de mama. Posteriormente, utilizou o modelo de riscos proporcionais de Cox (Farewell, 1982). Farewell & Sprott (1986) examinam o uso de tais modelos na inferência estatística. Goldman (1984) discute a análise de sobrevivência quando a cura é possível. Greenhouse & Wolfe (1984) estudam uma generalização do modelo de mistura padrão baseada na teoria de riscos competitivos.

Quando se utiliza a abordagem paramétrica nos modelos de mistura, é necessário assumir uma distribuição de probabilidade para o tempo até o evento de interesse dos indivíduos em risco. As funções densidade e de sobrevivência são obtidas da distribuição assumida, em que podem depender de um ou mais parâmetros como pode ser visto em Farewell (1982), Farewell & Sprott (1986) e Peng *et al.* (1998), entre outros. Maller & Zhou (1996) abordam o modelo de mistura padrão de uma perspectiva frequentista.

Considerando uma abordagem semiparamétrica para o modelo de mistura padrão, Kuk & Chen (1992) combinaram a formulação logística para a probabilidade de ocorrência do evento de interesse e assumem um modelo de riscos proporcionais para os indivíduos em risco. Para estimar os parâmetros do modelo proposto por eles foi utilizada simulação Monte Carlo e, desta forma, considerando uma generalização semiparamétrica para o modelo de Farewell (1982). Peng & Dear (2000) e Sy & Taylor (2000), propuseram usar o algoritmo EM para estimar os parâmetros.

Neste contexto, os modelos de sobrevivência de longa duração têm grande importância em análise de dados de sobrevivência e confiabilidade, e surgem em várias áreas, tais como medicina, finanças, criminologia e confiabilidade industrial. Por isso, diferentes métodos para ajustar tais modelos têm sido publicados na literatura. Diversos artigos têm abordado a questão dos dados de longa duração. Por exemplo, em confiabilidade industrial, o evento de interesse pode ser a falha de placas de circuito, devido a diferentes fatores de risco ou ao desgaste por uso, (Meeker & Escobar, 1998). Em dados financeiros, o evento de interesse pode ser o desligamento do cliente de um banco devido a várias causas (Hoggart & Griffin,

2001; Tong *et al.*, 2012). Em dados biomédicos, o evento de interesse pode ser a morte de um paciente submetido a certo tratamento, devido a diferentes causas competitivas ou à recorrência do tumor pela presença de um número desconhecido de células cancerígenas, como pode ser visto em Yakovlev & Tsodikov (1996),Chen *et al.* (1999) e Tsodikov *et al.* (2003) entre outros.

Recentemente, modelos mais complexos de longa duração como de Yakovlev & Tsodikov (1996), Chen *et al.* (1999), Ibrahim *et al.* (2001a), Rodrigues *et al.* (2009b), e outros, vêm sendo explorados com o objetivo de explicar melhor os mecanismos biológicos envolvidos.

Neste cenário, a metodologia proposta por Tsodikov *et al.* (2003) e Rodrigues *et al.* (2009a) entre outros, tem por objetivo unificar a análise de sobrevivência com o modelo clássico de Boag (1949) e Berkson & Gage (1952), e com os modelos mais recentes de longa duração Yakovlev & Tsodikov (1996), Chen *et al.* (1999). A unificação foi obtida através de uma composição da função geradora de probabilidade do número de causas de ocorrência do evento de interesse e da função de sobrevivência dos pacientes em risco (Tsodikov *et al.*, 2003). Neste contexto, a maioria dos modelos de longa duração fazem uso dessa proposta entre os quais podem ser citadas de Castro *et al.* (2009), Rodrigues *et al.* (2009b), Cancho *et al.* (2011) e Gu *et al.* (2011). Também é mostrado que a função geradora de longa duração formulada satisfaz a propriedade de riscos proporcionais se, e somente se, o número de causas relacionadas à ocorrência do evento de interesse segue uma distribuição de Poisson.

Na literatura estatística o modelo de mistura padrão é amplamente usado, no entanto possui algumas desvantagens que são discutidas em Chen *et al.* (1999). Estes autores fazem uso de um modelo com algumas vantagens em relação ao modelo de mistura padrão e esse modelo é conhecido na literatura como modelo de risco acumulado limitado (RAL) ou também denominado modelo de tempo de promoção.

Muitas extensões para o modelo de tempo de promoção foram propostas. Assim, baseado em estudos de câncer e de um ponto de vista paramétrico, tem-se o modelo desenvolvido por Hanin (2001), em que o número de riscos competitivos segue uma distribuição binomial negativa. Nesta mesma linha, recentemente, o modelo proposto por Rodrigues *et al.* (2011) leva em conta a sobredispersão e a subdispersão que usualmente está presente em dados discretos. Neste último artigo, a metodologia generaliza vários modelos inclusive o modelo de risco acumulado limitado.

Algumas extensões para o modelo de risco acumulado de um ponto de vista semipara-

métrico foram propostas por Ibrahim *et al.* (2001a), Kim *et al.* (2007) e outros. Em geral, os modelos semiparamétricos de longa duração constróem uma partição finita no eixo do tempo e assumem que a função de risco em cada subconjunto (intervalo) dessa partição é constante. A partição no eixo do tempo pode ser pré-especificada como em Ibrahim *et al.* (2001a) e Yin & Ibrahim (2005) ou ser considerada desconhecida (Kim *et al.*, 2007).

A motivação que fundamenta os modelos de partição é a de que pontos próximos em um espaço $\boldsymbol{\mathcal{X}}$, têm uma mesma distribuição local, isto é, pontos em uma mesma região têm uma mesma distribuição de probabilidade. A partir desta ideia, são construídas regiões sob $\boldsymbol{\mathcal{X}}$, de forma que as regiões são disjuntas entre si e a união delas é $\boldsymbol{\mathcal{X}}$. Assim, uma forma de obter uma partição de $\boldsymbol{\mathcal{X}}$ é usar uma estrutura de tesselação, por exemplo, a tesselação de Voronoi, tesselação por retângulos etc.

Considerando a ideia anterior, foram propostos modelos em que consideram a partição no espaço das covariáveis. Neste sentido, esses modelos de partição geralmente envolvem a partição considerando apenas uma covariável, isto é, o espaço preditor em dimensão 1, como exemplificado em Barry & Hartigan (1993), Stephens (1994) e em outros. Basicamente os autores anteriormente citados pesquisaram a análise de ponto de mudança. Existe também modelos para dados de sobrevivência que consideram a partição no espaço preditor dentro dos quais podemos citar Segal (1988) e Zhang & Singer (2010).

Existe ampla literatura em relação aos modelos de partição, por exemplo os modelos desenvolvidos por Quintana & Iglesias (2003), Hegarty & Barry (2008) e Muller & Quintana (2010) são baseados no modelo de partição produto (Hartigan, 1990; Barry & Hartigan, 1993). Outros modelos de partição podem ser vistos em Stephens (1994), Green (1995), Heikkinen (1998) e McCullagh & Yang (2008).

Um modelo que considera a partição no espaço das covariáveis, $\boldsymbol{\mathcal{X}}$, é o modelo de partição proposto por Holmes *et al.* (1999, 2005), esse modelo é conhecido na literatura como modelo de partição bayesiana (MPB). Neste sentido, a fim de obter uma partição em $\boldsymbol{\mathcal{X}}$ o modelo MPB faz uso de uma tesselação, como a tesselação de Voronoi. O modelo MPB foi inicialmente proposto para abordar problemas de classificação e regressão, porém extensões para modelar dados discretos foram propostos por Denison & Holmes (2001). Além disso, uma característica principal do modelo MPB é que assume independência entre as regiões de partição do espaço preditor.

Recentemente, têm sido desenvolvidas pesquisas envolvendo extensões bayesianas para modelos clássicos.

1.2 Objetivos do trabalho

Para a análise de dados de longa duração, isto é, quando se admite uma porcentagem de não ocorrência do evento de interesse na população, é considerado um modelo de longa duração em que o número de riscos competitivos segue uma distribuição de série de potências. Esse modelo de longa duração será chamado de modelo de série de potências com fração de cura. Neste sentido, os modelos de risco acumulado e de mistura padrão são casos particulares do modelo de série de potências com fração de cura.

Na presença de covariáveis, usualmente os modelos de longa duração relacionam a fração de cura com as variáveis preditoras por meio de uma função de ligação, por exemplo, no modelo de mistura padrão em geral consideram uma função de ligação logito (Kuk & Chen, 1992; Peng & Dear, 2000) e no modelo de risco acumulado a função de ligação logarítmica (Chen *et al.*, 1999). Neste sentido, geralmente os modelos de longa duração assumem a linearidade das covariáveis, porém isto não sempre é real. Além disso, não todas as covariáveis consideradas para ajuste do modelo tem um efeito na fração de cura. Em seguida, também existe a possibilidade que um subconjunto de amostra tenham um comportamento homogêneo, por exemplo se indivíduos com câncer são submetidos a um processo de tratamento (e.g., quimioterapia) e ao final do processo existem grupos de indivíduos que respondam de maneira similar ao tratamento.

Neste contexto, um objetivo deste trabalho é propor um modelo de longa duração que leve em conta a não linearidade dos dados. Assim, propõe-se uma extensão local do modelo de longa duração de série de potências, a extensão local é baseada no modelo de partição bayesiana. Em seguida, levando em conta a partição no espaço preditor tem-se que a fração de cura depende das covariáveis de forma local e desta forma a fração de cura captura os efeitos locais. A fim de obter uma partição no espaço das covariáveis, $\boldsymbol{\mathcal{X}}$ neste trabalho foi adotado a tesselação por hiperplanos ortogonais paralelos aos eixos.

Um segundo objetivo deste trabalho é considerar a seleção de covariáveis na extensão local do modelo de série de potências com fração de cura. Neste sentido, uma vez que foi adotado a tesselação por hiperplanos ortogonais aos eixos tem-se que a seleção de variáveis preditoras pode ser feita. Assim, se uma covariável não é informativa no modelo então essa variável preditora não será dividida e portanto pode-se afirmar que essa variável preditora não tem efeito na fração de cura.

Um terceiro objetivo deste trabalho é encontrar agrupamentos nos dados que tenham um comportamento similar (homogêneo) em relação a probabilidade de ser curado.

Neste trabalho, é considerado uma abordagem bayesiana para o modelo proposto. Assim, para obter amostras da distribuição *a posteriori* do modelo proposto é considerada uma estrategia computacional baseado em métodos de simulação Monte Carlo via cadeias de Markov (MCMC).

1.3 Organização do trabalho

Este trabalho está organizado da seguinte forma: no Capítulo 2, é apresentado o modelo de série de potências com fração de cura baseado na metodologia proposta por Tsodikov *et al.* (2003) e Rodrigues *et al.* (2009a). Esse modelo é aplicado a dois conjunto de dados reais. Além disso, foi considerado uma abordagem frequentista para obter estimativas dos parâmetros. Alguns resultados deste capítulo foram condensados no artigo Gonzales *et al.* (2013).

No Capítulo 3 é apresentada o modelo de partição bayesiana. Não obstante, neste trabalho é considerado covariáveis quantitativas e qualitativas (com mais de duas categorias) e assim a abordagem proposta é uma extensão da metodologia proposta por Holmes *et al.* (1999, 2005). Neste sentido, a fim de explorar a distribuição *a posteriori* do modelo de partição é proposto um algoritmo MCMC que leva em conta a natureza da variável preditora.

No Capítulo 4, propõe-se a extensão local para o modelo de série de potências com fração de cura, considerando o modelo de partição bayesiana apresentado no Capítulo 3. O modelo proposto foi aplicado a dois conjuntos de dados reais. Resultaram deste capítulo os relatórios técnicos Gonzales *et al.* (2012), Tomazella *et al.* (2012) e Tomazella *et al.* (2013) e um artigo Louzada *et al.* (2014)

No Capítulo 5, encontram-se as considerações finais e as propostas futuras do trabalho.

Capítulo 2

Modelo de série de potências com fração de cura

Neste capítulo, é apresentada uma introdução aos modelos de longa duração conhecidos também como modelos de sobrevivência com fração de cura. Grande parte dos modelos de longa duração é aplicada e desenvolvida em estudos de câncer e epidemiológicos. Apesar de haver vasta literatura nesta área, os métodos estatísticos para análise de dados desse tipo ainda não estão disseminados e o assunto continua sendo alvo de muitas discussões. Neste estudo, são apresentados dois exemplos de aplicação relacionados com estudos de câncer.

Exemplo 2.1. Considere-se o conjunto de dados de leucemia que está disponível no pacote dynpred em R (Putter, 2011), em que todos os indivíduos tiveram um transplante de médula óssea alogênico de um irmão doador HLA (Human Leukocyte Antigens - Antígenos de Histocompatibilidade Humano) idêntico, entre 1985 e 1998. Nesse conjunto de dados tem-se 1764 pacientes com um quadro de leucemia linfóide aguda leucemia (ALL) em que as covariáveis foram observadas no tempo do transplante. As frequências para cada covariável são apresentadas na Tabela 2.1.

Na Figura 2.1 (esquerda), apresenta-se a estimativa de Kaplan-Meier (K-M) da função de sobrevivência estratificada pela variável idade dos pacientes, em que é possível notar que no caso dos pacientes que têm idades iguais ou menores que 20 anos, a curva de K-M se estabiliza acima de 0,4. Assim, devido a esse comportamento os modelos que não levam em conta uma proporção de curados podem não ser adequados para o analise de estes dados. Um comportamento similar é observado na estimativa de função de sobrevivência

Tabela 2.1: Frequencias das covariáveis para o conjunto de dados de leucemia LLA.

Covariável	Categorias	Frequência
x_1: Ano do transplante	1985-1989	561
	1990-1994	682
	1995-1998	521
x_2: Idade do paciente	≤ 20	551
	20-40	1213
x_3: Profilaxia	Sim	1353
	Não	411
x_4: Incompatibilidade doador-receptor	Incompatibilidade de gênero	433
	Compatibilidade de gênero	1331

de K-M para o grupo de pacientes com idade entre os 20 e 40 anos. Observa-se que as curvas de K-M para o grupo de indivíduos que receberam o transplante de médula óssea entre 1990-1994 e 1995-1998 são próximos e intuitivamente os pacientes que receberam o transplante de médula óssea nesses anos têm um comportamento similar e podem ser combinados em um único grupo.

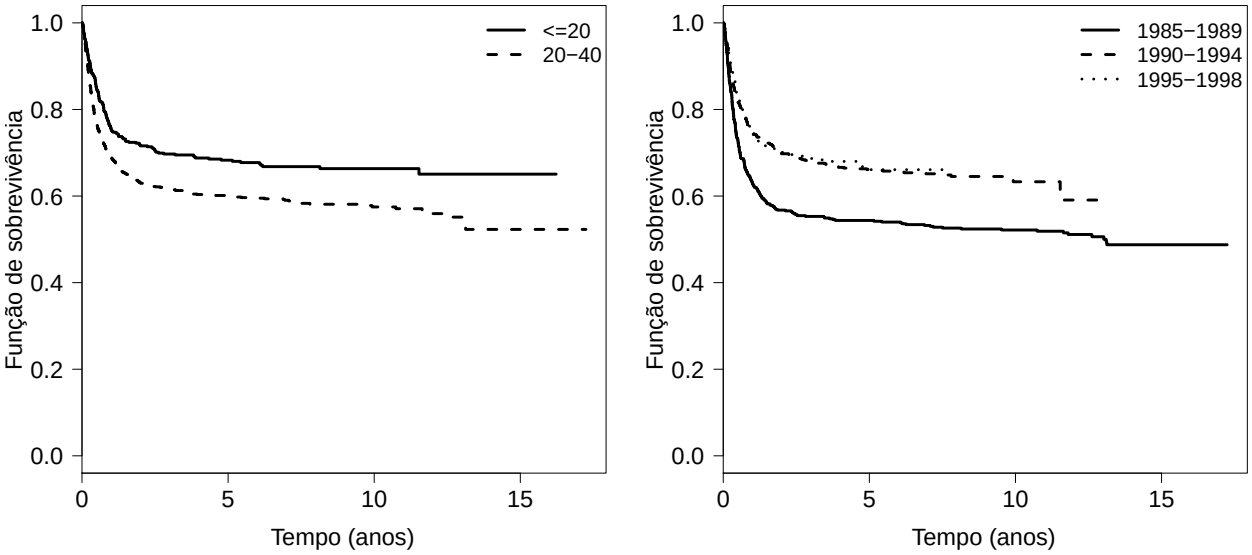


Figura 2.1: Estimativa de K-M da função de sobrevivência para os dados de leucemia aguda, considerando-se as covariáveis idade (painel esquerdo) e ano de transplante (painel direito) .

Na Figura 2.2 é apresentada a estimativa da função de risco acumulado baseado no estimador proposto por Nelson-Aalen (Nelson, 1972; Aalen, 1978), assim pode-se notar que a probabilidade de acontecer um relapso ou morte nos primeiros dois anos é alto e, desta forma, o risco é maior, nesse intervalo de tempo. Porém, o risco começa a decrescer depois dos dois primeiros anos, e intuitivamente tende a se estabilizar, o que nos leva a suspeitar que existe uma proporção de pacientes que não morrem ou não experimentam o relapso e assim podem ser considerados como pacientes "curados".

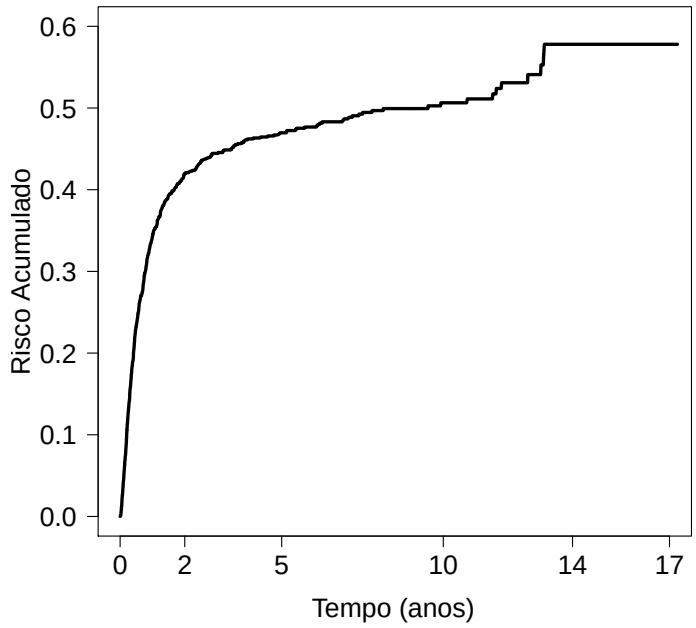


Figura 2.2: Estimativa da função de risco acumulado para os dados de leucemia aguda no Exemplo 2.1.

Exemplo 2.2. Kirkwood *et al.* (2000) e Ibrahim *et al.* (2001b) consideraram um conjunto de dados de um estudo de melanoma cutâneo (um tipo de câncer) com o objetivo de avaliar (pós-operatório) a eficácia da aplicação de uma dosagem alta de interferon alfa-2b como forma de prevenir a recorrência de câncer. Os pacientes foram incluídos no estudo entre 1991 e 1995, tendo sido acompanhados até 1998. A variável resposta T representa o tempo até a morte de paciente ou tempo de censura. Nesta amostra, tem-se n =417 pacientes, com 56% de observações censuradas. As variáveis incluem y: tempo (em anos); x_1: tipo de tratamento (0: sem tratamento; 1:interferon) ; x_2: idade; x_3: categoria do nódulo (1,2,3,4); x_4: sexo (0: masculino; 1: feminino); x_5: capacidade funcional (0: ativo; 1: outras) e x_6: espessura do tumor.

A Figura 2.3 mostra a estimativa de Kaplan-Meier (K-M) da função de sobrevivência para este conjunto de dados, onde se pode observar que, após um determinado tempo, a curva se estabiliza não havendo mais falhas. Isto sugere que os indivíduos censurados no final do experimento possam ser imunes ao risco em questão ou foram curados durante o experimento. Assim, utilizar as técnicas usuais em análise de sobrevivência para o analise dos dados descritos anteriormente podem não ser adequadas isto pelo fato que não incorporam a fração de cura.

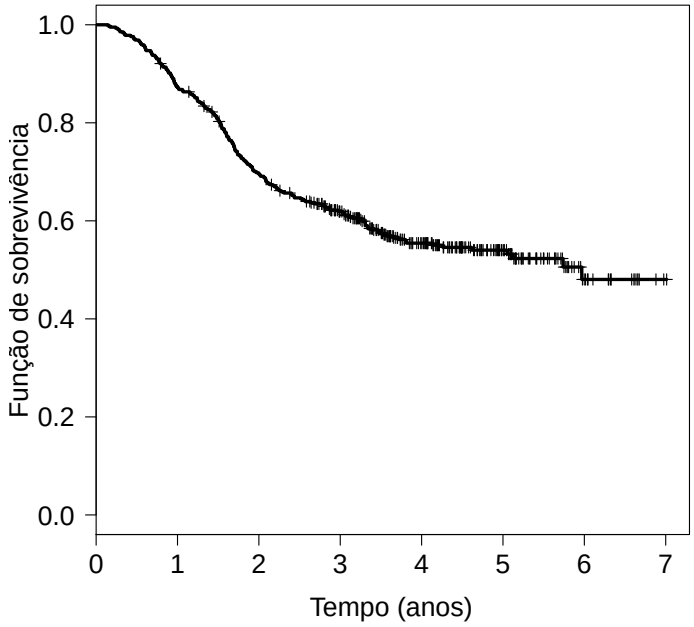


Figura 2.3: Estimativa de K-M da função de sobrevivência para os dados de melanoma no Exemplo 2.2.

Os exemplos anteriores mostram evidências que existem um grupo de individuos que não apresentaram o evento de interesse (morte ou relapso da doença) e desta forma na Seção 2.1 será apresentada uma modelagem estatística capaz de levar em conta essa proporção de curados na população.

2.1 Modelagem de fração de cura

Seja N uma variável aleatória (v.a.) que representa o número de causas ou riscos para um particular evento de interesse e assuma-se que tenha distribuição de probabilidade

$$p_{n^*} = P[N = n^*], \quad n^* = 0, 1, 2, \ldots,$$

em que N é uma variável aleatória latente. Condicionado a $N = n^*$, sejam Z_v, $v = 1, \ldots, n^*$, variáveis aleatórias contínuas não negativas e independentes, com função de distribuição $F(t) = 1 - S(t)$, sendo que N é independente de Z_v, em que Z_v representa o tempo de ocorrência de um particular evento de interesse, devido à v-ésima causa ou risco.

O tempo de ocorrência do evento de interesse é definido como

$$T = \min\{Z_1, Z_2 \ldots, Z_N\}, \tag{2.1}$$

se $N \geq 1$. No caso que $N = 0$ tem-se que $T = \infty$ ($Z_0 = \infty$), e desta forma existe uma proporção p_0 da população não sujeita à ocorrência do evento de interesse. As variáveis aleatórias Z_v são variáveis latentes e T é uma variável aleatória observável ou censurada.

A função de sobrevivência da variável aleatória T [cf. (2.1)] é chamada de função de sobrevivência da população e de acordo com Tsodikov *et al.* (2003) e Rodrigues *et al.* (2009a) é dada por

$$S_{pop}(t) = P[T > t] = A_N(S(t)) = \sum_{n^*=0}^{\infty} p_{n^*} \{S(t)\}^{n^*}, \tag{2.2}$$

em que $A_N(\cdot)$ é a função geradora de probabilidade da variável aleatória N, e é convergente para valores de $s = S(t) \in [0, 1]$.

A proporção de não ocorrência do evento de interesse na população, p_0, é dada por

$$\lim_{t\to\infty} S_{pop}(t) = p_0, \tag{2.3}$$

e levando em conta isto tem-se que a função de sobrevivência $S_{pop}(t)$ dada em (2.2) não é uma função de sobrevivência própria , isto é, $0 < S_{pop}(\infty) < 1$. As funções de densidade e de risco associadas à função de sobrevivência de longa duração $S_{pop}(t)$ são dadas, respectivamente por

$$f_{pop}(t) = f(t)\frac{\mathrm{d}A_N(s)}{\mathrm{d}s}\,|_{s=S(t)} \tag{2.4}$$

e

$$h_{pop}(t) = \frac{f_{pop}(t)}{S_{pop}(t)} = f(t)\frac{\frac{\mathrm{d}A_N(s)}{\mathrm{d}s}\,|_{s=S(t)}}{S_{pop}(t)}. \tag{2.5}$$

Observe-se que em (2.5), não ocorre necessariamente a propriedade de riscos proporcionais considerando que

$$\frac{\frac{\mathrm{d}A_N(s)}{\mathrm{d}s}\,|_{s=S(t)}}{S_{pop}(t)}$$

pode depender de t. No entanto, uma exceção acontece quando a variável aleatória N tem distribuição de Poisson com parâmetro $\theta > 0$, isto é $N \sim \text{Poi}(\theta)$.

Exemplos de funções geradoras de probabilidade, associadas a algumas distribuições de probabilidade para $0 \leq s \leq 1$, são Bernoulli, binomial, geométrica, Poisson entre outras.

Neste trabalho, considera-se que o número de causas competitivas, N, segue uma distribuição da família de série de potências (Johnson *et al.*, 2005). Uma vantagem desse modelo é que é flexível, pois inclui casos particulares, tais como a distribuição binomial, Poisson, binomial negativa e logarítmica.

Assim, seja N uma variável aleatória que segue a distribuição de série de potências (DSP) com distribuição de probabilidade dada por

$$P[N = n^*] = \frac{a_{n^*}\theta^{n^*}}{\eta(\theta)}, \quad n^* = 0, 1, 2, \ldots, \quad \theta > 0, \tag{2.6}$$

em que $a_{n^*} \geq 0$ e $\eta(\theta) = \sum_{n^*=0}^{\infty} a_{n^*}\theta^{n^*} < +\infty$. Em (2.6), θ é chamado de parâmetro de potência e $\eta(\cdot)$ é conhecido como função da série.

A função geradora de probabilidade de N neste caso, é dada por

$$A_N(s) = \frac{\eta(\theta s)}{\eta(\theta)}, \quad 0 \leq s \leq 1, \tag{2.7}$$

sendo que a média e variância de N são dadas, respectivamente, por

$$\mathrm{E}[N] = \mu = \theta\frac{\mathrm{d}}{\mathrm{d}\theta}\log(\eta(\theta))$$

e

$$\mathrm{Var}[N] = \theta^2\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log(\eta(\theta)) + \mu.$$

As distribuições de probabilidade Bernoulli, Poisson e geométrica são casos particulares da família de distribuição de série de potências por considerar diferentes funções de série $\eta(\cdot)$. Assim, por exemplo, se for considerado como função de série $\eta(\theta) = (1+\theta)^K$ tem-se que a distribuição de série potencias se reduz à distribuição binomial com parâmetros K e $\theta/(1+\theta)$, em que K é um inteiro positivo e θ um número real positivo. Logo, Kosambi (1949) e Noack (1950) chegaram aos resultados seguintes que são apresentados na Tabela 2.2 em que o parâmetro τ da distribuição binomial negativa é um inteiro positivo.

Tabela 2.2: Distribuição de N para diferentes funções de série $\eta(\cdot)$

Distribuição	Suporte	Θ	a_{n^*}	a_0	$\eta(\theta)$
Binomial $\mathrm{Bi}\left(K, \frac{\theta}{1+\theta}\right)$	$\{0,1,\ldots,K\}$	$(0,\infty)$	$\binom{K}{n^*}$	1	$(1+\theta)^K$
Poisson $\mathrm{Poi}(\theta)$	$\{0,1,2,\ldots\}$	$(0,\infty)$	$\frac{1}{n^*!}$	1	e^θ
Binomial negativa $\mathrm{Bn}(\tau,\theta)$	$\{0,1,2,\ldots\}$	$(0,1)$	$\binom{\tau+n^*-1}{\tau-1}$	1	$(1-\theta)^{-\tau}$
Logarítmica $\mathrm{Lg}(\theta)$	$\{0,1,2,\ldots\}$	$(0,1)$	$\frac{1}{(n^*+1)}$	1	$\frac{-\log(1-\theta)}{\theta}$

A função de sobrevivência da população para a distribuição de série de potências é dada por

$$S_{pop}(t) = A_N(S(t)) = \frac{\eta(\theta S(t))}{\eta(\theta)}. \tag{2.8}$$

A fração de cura p_0 é dada, por

$$p_0 = \lim_{t\to\infty} S_{pop}(t) = \frac{\eta(\theta S(\infty))}{\eta(\theta)} = \frac{a_0}{\eta(\theta)} < 1. \tag{2.9}$$

As funções de densidade e de risco associadas à função de sobrevivência de longa duração dada em (2.8) são dadas, respectivamente, por

$$f_{pop}(t) = \frac{\eta'(\theta S(t))}{\eta(\theta)}\theta f(t) \quad \text{e} \quad h_{pop}(t) = \frac{\eta'(\theta S(t))}{\eta(\theta S(t))}\theta f(t), \tag{2.10}$$

em que $f(t) = \mathrm{d}F(t)/\mathrm{d}t$. Observa-se que $f_{pop}(\cdot)$ e $h_{pop}(\cdot)$ são funções impróprias, isto se deve ao fato que $S_{pop}(\cdot)$ não é uma função de sobrevivência própria.

2.2 Casos particulares

Vários modelos de longa duração são casos particulares do modelo de série de potências com fração de cura (MLDDSP) e são apresentados nesta seção.

2.2.1 Modelo de longa duração binomial (MLDBi)

Nesta seção descreve-se o modelo de longa duração binomial, isto é, quando o número N de riscos latentes é assumido com distribuição binomial. Uma motivação biológica deste

modelo foi proposta por Gail *et al.* (1980), no qual, supondo-se que existe um número K de potenciais lugares para a mutação de tumores localizados numa região do corpo de um indivíduo afetada por uma doença (e.g., câncer) tem-se que N ($N \leq K$) lugares chegam a sofrer mutações.

Adota-se uma reparametrização do modelo binomial adotando a transformação $\theta^* = \theta/(1+\theta)$. Assim , seja N uma variável aleatória que representa o número de causas competitivas latentes necessárias para a ocorrência de um determinado evento de interesse, que segue uma distribuição binomial com função de probabilidade de massa dada por

$$P[N = n^*] = \binom{K}{n^*} \theta^{*n^*}(1-\theta^*)^{K-n^*}, \quad n^* = 0, 1, \ldots, K, \quad 0 < \theta^* < 1, \tag{2.11}$$

com $\mathrm{E}[N] = K\theta^*$ e $\mathrm{Var}[N] = K\theta^*(1-\theta^*)$, em que K é um número inteiro positivo. A função geradora de probabilidade de N é dada por

$$A_N(s) = (1 - \theta + \theta^* s)^K, \quad 0 \leq s \leq 1. \tag{2.12}$$

A função de sobrevivência de longa duração é dada por

$$S_{pop}(t) = A_N(S(t)) = \{1 - \theta^* + \theta^* S(t)\}^K, \tag{2.13}$$

sendo que a fração de cura p_0, é dada por

$$p_0 = \lim_{t\to\infty} S_{pop}(t) = (1-\theta^*)^K > 0. \tag{2.14}$$

As funções de densidade e de risco impróprias associadas à função de sobrevivência de longa duração binomial são dadas respectivamente por

$$f_{pop}(t) = K\theta^* f(t)\{1 - \theta^* + \theta^* S(t)\}^{(K-1)} \quad \text{e} \quad h_{pop}(t) = \frac{K\theta^* f(t)}{\{1 - \theta^* + \theta^* S(t)\}}, \tag{2.15}$$

em que $f(t) = -\mathrm{d}S(t)/\mathrm{d}t$ é uma função de densidade própria.

Um caso particular do modelo binomial de longa duração é o modelo de mistura padrão (MLDBer), isto é, quando $K = 1$. Observa-se que, no caso em que K cresce, a fração de cura p_0 decresce. Ao longo deste trabalho assumimos que o parâmetro K é fixo, embora em um cenário mais realista, poderia ser considerada uma distribuição de probabilidade para K, porém a complexidade computacional é maior (Cooner *et al.*, 2007).

Em perspectiva bayesiana, Chen *et al.* (1999) mostram algumas desvantagens do modelo de mistura padrão. Especificamente, relacionando-se as covariáveis com o parâmetro θ^* via um modelo de regressão binomial e se é considerada uma distribuição *a priori* imprópria para os coeficientes da regressão em θ^*, tem-se que a distribuição *a posteriori* dos parâmetros do modelo é imprópria.

2.2.2 Modelo de longa duração Poisson (MLDPoi)

Se for considerado como função de série $\eta(\theta) = e^{\theta}$ e $a_{n^*} = 1/n^*!$, tem-se que o número N de causas do evento de interesse tem a distribuição de probabilidade de Poisson. Assim, considerando a função geradora de probabilidade da distribuição de Poisson $A_N(s) = \exp\{\theta(1-s)\}$, obtém-se as funções de sobrevivência, densidade e de risco da população, dadas respectivamente por

$$S_{pop}(t) = A_N(S(t)) = \exp\left\{-\theta F(t)\right\}, \tag{2.16}$$

$$f_{pop}(t) = -\frac{\mathrm{d}S_{pop}(t)}{\mathrm{d}t} = \theta f(t) \exp\left\{-\theta F(t)\right\} \tag{2.17}$$

e

$$h_{pop}(t) = \theta f(t). \tag{2.18}$$

Assim, de (2.16) tem-se a fração de cura dada por $p_0 = \lim_{t\to\infty} S_{pop}(t) = \exp(-\theta)$. O modelo de longa duração definido em (2.16) é conhecido como modelo de risco acumulado limitado (RAL) ou modelo de tempo de promoção.

As funções de sobrevivência, densidade e de risco para a população em risco são dadas respectivamente, por

$$S^*(t) = \frac{\exp(-\theta F(t)) - \exp(-\theta)}{1 - \exp(-\theta)}, \tag{2.19}$$

$$f^*(t) = \left(\frac{\exp(-\theta F(t))}{1 - \exp(-\theta)}\right) \theta f(t)$$

e

$$h^*(t) = \left(\frac{\exp(-\theta F(t))}{\exp(-\theta F(t)) - \exp(\theta)}\right) h_{pop}(t).$$

A relação matemática do modelo de tempo de promoção com o modelo de mistura padrão é dada por

$$S_{pop}(t) = \exp(-\theta) + (1 - \exp(-\theta))S^*(t), \tag{2.20}$$

em que $S^*(t)$ é dada por (2.19). Desse modo, $S_{pop}(t)$ tem a forma do modelo de mistura padrão com fração de cura $p_0 = 1 - \exp(-\theta)$.

Uma característica do modelo de tempo de promoção é que a função de risco da população dada em (2.18), tem a propriedade de riscos proporcionais.

O modelo RAL dada em (2.16) foi proposto por Yakovlev & Tsodikov (1996) baseado em considerações biológicas para a recorrência de um tumor. Além disso, tais autores assumiram uma abordagem paramétrica para o modelo de tempo de promoção. Para as estimativas dos parâmetros basearam-se no método de máxima verossimilhança. Uma abordagem bayesiana foi proposta por Chen *et al.* (1999).

2.2.3 Modelo de longa duração binomial negativa (MLDBn)

Nesta seção, a modelagem proposta baseia-se em que o número de causas competitivas N segue uma distribuição binomial negativa. Assim, por adotar que a função de série é dada por $\eta(\theta) = (1-\theta)^{-\tau}$ e $a_{n^*} = \binom{\tau+n^*-1}{\tau-1}$ tem-se que N tem função de probabilidade definida por

$$P[N = n^*] = \binom{\tau + n^* - 1}{\tau - 1}\theta^{n^*}(1-\theta)^{\tau}, \quad n^* = 0, 1, 2, \ldots, \quad 0 < \theta < 1, \tag{2.21}$$

em que τ é um inteiro positivo. A média e a variância de N são respectivamente

$$\mathrm{E}[N] = \tau\theta/(1-\theta) \quad \text{e} \quad \mathrm{Var}[N] = \tau\theta/(1-\theta)^2. \tag{2.22}$$

A função geradora de probabilidades é dada por

$$A_N(s) = \sum_{n^*=0}^{\infty} p_{n^*}s^{n^*} = \left(\frac{1-\theta}{1-\theta s}\right)^{\tau}, \quad 0 \leq s \leq 1. \tag{2.23}$$

Assim, a função de sobrevivência de longa duração para o modelo binomial negativa é dada por

$$S_{pop}(t) = A_N(S(t)) = \left(\frac{1-\theta}{1-\theta S(t)}\right)^{\tau}. \tag{2.24}$$

A função densidade imprópria do modelo dada em (2.24) é

$$f_{pop}(t) = -\frac{\mathrm{d}S_{pop}(t)}{\mathrm{d}t} = \frac{\tau\theta(1-\theta)^{\tau}f(t)}{(1-\theta S(t))^{\tau+1}}, \tag{2.25}$$

em que $f(t) = -\mathrm{d}S(t)/\mathrm{d}t$. Além disso, a função de risco correspondente é dada por

$$h_{pop}(t) = \frac{\tau\theta f(t)}{(1-\theta S(t))}. \tag{2.26}$$

De (2.24) tem-se que a fração de cura é dada por

$$p_0 = \lim_{t\to\infty} S_{pop}(t) = (1-\theta)^{\tau}.$$

No caso em que $\tau = 1$ tem-se como caso particular a distribuição geométrica. Uma motivação biológica é encontrado em estudos clínicos de câncer. Por exemplo, um indivíduo com exposição a dano genético faz com que ela produza N células mutantes antes que o sistema imune seja ativado (Moolgavkar *et al.*, 1990). Seguindo Cooner *et al.* (2007), cada nova célula mutada produz uma resposta efetiva do sistema imune capaz de destruir a última célula mutante com probabilidade $1-\theta$, então N segue uma distribuição geométrica com parâmetro θ.

2.2.4 Modelo de longa duração logarítmica (MLDLg)

No caso em que N segue a distribuição logarítmica, a função de série é dada por

$$\eta(\theta) = -\log(1-\theta),$$

e $a_{n^*} = 1/n^*$. Assim, a função de distribuição para N é definida sendo

$$P[N = n^*] = \frac{\theta^{n^*}}{-n^* \log(1-\theta)}, \quad n^* = 1, 2, \ldots, \quad 0 < \theta < 1.$$

Porém a fração de cura é definida quando $P[N = 0]$, assim desloca-se o domínio da distribuição logarítmica em zero e a função de probabilidade fica definida como

$$P[N = n^*] = \frac{\theta^{n^*+1}}{-(n^*+1)\log(1-\theta)} \quad n^* = 0, 1, \ldots, \quad 0 < \theta < 1, \tag{2.27}$$

em que a função de série é dada por

$$\eta(\theta) = -\frac{\log(1-\theta)}{\theta},$$

em que $a_{n^*} = 1/(n^*+1)$. A média e variância para N são dadas respectivamente por

$$\mathrm{E}[N] = \frac{a^*\theta}{1-\theta} - 1 \quad \text{e} \quad \mathrm{Var}[N] = \frac{a^*\theta(1-a^*\theta)}{(1-\theta)^2},$$

em que $a^* = -1/\log(1-\theta)$. A função geradora da probabilidade para a distribuição logarítmica definida em (2.27) é dada por

$$A_N(s) = \frac{\log(1-\theta s)}{s\log(1-\theta)}.$$

Sendo que o número de causas latentes até o evento de interesse segue a distribuição logarítmica, a função de sobrevivência de longa duração é dada por

$$S_{pop}(t) = A_N(S(t)) = \frac{\log(1-\theta S(t))}{S(t)\log(1-\theta)}. \tag{2.28}$$

A função densidade correspondente do modelo (2.28) é

$$f_{pop}(t) = -\frac{dS_{pop}(t)}{dt} = -f(t)\frac{\theta S(t) + (1-\theta S(t))\log(1-\theta S(t))}{S^2(t)(1-\theta S(t))\log(1-\theta)},$$

em que $f(t) = -\mathrm{d}S(t)/\mathrm{d}t$. Além disso, a função de risco correspondente é dada por

$$h_{pop}(t) = -f(t)\log(1-\theta S(t))\frac{\theta S(t) + (1-\theta S(t))\log(1-\theta S(t))}{S(t)(1-\theta S(t))}.$$

De (2.28) tem-se que a fração de cura é dada por

$$p_0 = \lim_{t\to\infty} S_{pop}(t) = -\frac{\theta}{\log(1-\theta)}.$$

Na Tabela 2.3 são apresentados a função de sobrevivência de longa duração, a densidade imprópria e fração de cura correspondentes aos modelos estudados nas seções anteriores.

Tabela 2.3: Função de sobrevivência $S_{pop}(t)$, função de densidade $f_{pop}(t)$ e fração de cura para diferentes distribuições do número de causas latentes, N.

Distribuição	$S_{pop}(t)$	$f_{pop}(t)$	p_0
$\mathrm{Bi}(K,\theta^*)$	$(1-\theta^*+\theta^*S(t))^K$	$K\theta^* f(t)(1-\theta^*+\theta^*S(t))^{K-1}$	$(1-\theta^*)^K$
$\mathrm{Poi}(\theta)$	$\exp(-\theta F(t))$	$\theta f(t)\exp(-\theta F(t))$	$e^{-\theta}$
$\mathrm{Bn}(\tau,\theta)$	$\left(\frac{1-\theta}{1-\theta S(t)}\right)^\tau$	$\frac{\tau\theta(1-\theta)^\tau f(t)}{(1-\theta S(t))^{\tau+1}}$	$(1-\theta)^\tau$
$\mathrm{Lg}(\theta)$	$\frac{\log(1-\theta S(t))}{S(t)\log(1-\theta)}$	$-f(t)\frac{\theta S(t)+(1-\theta S(t))\log(1-\theta S(t))}{S^2(t)(1-\theta S(t))\log(1-\theta)}$	$\frac{-\theta}{\log(1-\theta)}$

2.3 Inferência

Considera-se o cenário em que o tempo definido em (2.1) não é completamente observável e está sujeito a censura à direita. Seja C_i o tempo da censura para o i ésima unidade amostral e $Y_i = \min\{T_i, C_i\}$ o tempo observado e δ_i a variável indicadora de censura em que $\delta_i = 1$ se $Y_i = T_i$, e $\delta_i = 0$, caso contrário, $i = 1, 2, \ldots, n$.

Seja $\boldsymbol{\gamma}$ o vetor de parâmetros da distribuição do tempo não observado dada em (2.1). Na presença de covariáveis, seja $\boldsymbol{x}_i^\top = (x_{i0}, x_{i1}, \ldots, x_{ip})$ que denota o vetor de covariáveis associado ao i ésimo indivíduo em que inclui um intercepto ($x_{i0} = 1$) e seja $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ o vetor de coeficientes da regressão. Foram introduzidas covariáveis no parâmetro θ_i por meio de uma função de ligação $g(.)$. Assim para o modelo binomial, binomial negativa e logarítmica adotou-se a função de ligação

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \quad i = 1, 2, \ldots, n,$$

e, para o modelo Poisson, considera-se

$$\log(\theta_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \quad i = 1, 2, \ldots, n.$$

Foram denotados os dados completos por $\mathcal{D}_c = (n, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{N})$, em que $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^\top$, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$, a matriz das covariáveis de ordem $n \times (p+1)$ e $\boldsymbol{N} = (N_1, \ldots, N_n)^\top$ o vetor de variáveis latentes. A função de verossimilhança com dados completos supondo censura não informativa, é dada por

$$L(\boldsymbol{\vartheta}; \mathcal{D}_c) = \prod_{i=1}^{n} \{S(y_i|\boldsymbol{\gamma})\}^{N_i - \delta_i} \{N_i f(y_i|\boldsymbol{\gamma})\}^{\delta_i} p_{n_i^*}(\theta_i), \tag{2.29}$$

em que $\boldsymbol{\vartheta} = (\boldsymbol{\gamma}^\top, \boldsymbol{\beta}^\top)^\top$. O vetor latente $\boldsymbol{N}$ é não observável. Fazendo-se o somatório ao longo do vetor $\boldsymbol{N}$ em (2.29), obtém-se a função de verossimilhança baseada nos dados observados $\mathcal{D} = (n, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\delta})$ que é dada por

$$L(\boldsymbol{\vartheta}; \mathcal{D}) = \sum_{\boldsymbol{N}} L(\boldsymbol{\vartheta}; \boldsymbol{\mathcal{D}}_c). \tag{2.30}$$

Neste cenário, considerando a metodologia proposta por Rodrigues *et al.* (2009a) tem-se que a função de verossimilhança para os dados observados é dada por

$$L(\boldsymbol{\vartheta}|\mathcal{D}) \propto \prod_{i=1}^{n} \{f_{pop}(y_i|\boldsymbol{\vartheta})\}^{\delta_i} \{S_{pop}(y_i|\boldsymbol{\vartheta})\}^{1-\delta_i}, \tag{2.31}$$

em que $f_{pop}(.)$ e $S_{pop}(.)$ são as funções dadas para os modelos definidos na Tabela 2.3. Foi adotada a distribuição Weibull para o tempo até a ocorrência Z [cf. (2.1)] sendo que a função densidade e de distribuição são dadas respectivamente por

$$f(y;\boldsymbol{\gamma}) = \alpha y^{\alpha-1}\exp(\lambda - y^{\alpha}e^{\lambda}) \quad \text{e} \quad F(y;\boldsymbol{\gamma}) = 1-\exp(-y^{\alpha}e^{\lambda}), \tag{2.32}$$

em que $\boldsymbol{\gamma} = (\alpha,\lambda)^{\top}$, tal que $\alpha > 0$ and $\lambda \in \mathbb{R}$.

De um ponto de vista frequentista, as estimativas para os parâmetros $(\boldsymbol{\gamma},\boldsymbol{\beta})$ são obtidas pelo método de máxima verossimilhança. Para este fim maximiza-se o logaritmo da função de verossimilhança $\ell(\boldsymbol{\vartheta};\mathcal{D}) = \log L(\boldsymbol{\vartheta};\mathcal{D})$ usando métodos numéricos de maximização. A programação computacional foi feita no software **R** (R Core Team, 2013) usando a função `optim`.

Sob certas condições de regularidade pode-se mostrar que a distribuição assintótica do estimador de máxima verossimilhança, $\widehat{\boldsymbol{\vartheta}}$, segue uma distribuição normal multivariada com vetor de médias $(\boldsymbol{\beta},\boldsymbol{\gamma})$ e matriz de covariância $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\vartheta}})$, que pode ser estimado por

$$\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\vartheta}}) = \left\{-\frac{\partial^2\ell(\boldsymbol{\vartheta};\mathcal{D})}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^{\top}}\right\}^{-1}, \tag{2.33}$$

avaliado em $\boldsymbol{\vartheta} = \widehat{\boldsymbol{\vartheta}}$ sendo que as primeiras e segundas derivadas do logaritmo da função de verossimilhança, $\ell(\boldsymbol{\vartheta};\mathcal{D})$, são obtidas numericamente.

Para comparar o ajuste dos modelos propostos, utilizam-se o critério de informação de Akaike (AIC) e o critério de informação bayesiano (BIC) dados por AIC $= -2\ell(\widehat{\boldsymbol{\vartheta}}) + 2\#(\widehat{\boldsymbol{\vartheta}})$ e BIC $= -2\ell(\widehat{\boldsymbol{\vartheta}}) + \#(\widehat{\boldsymbol{\vartheta}})\log(n)$, em que $\#(\widehat{\boldsymbol{\vartheta}})$ é o número de parâmetros do modelo ajustado. Para escolher um modelo dentro de um conjunto finito de candidatos de modelos, selecionou-se o modelo que tem menor valor para ambos critérios.

2.4 Aplicação

Para mostrar a metologia desenvolvida neste capítulo são considerados os conjunto de dados descritos no Exemplo 2.1 e 2.2.

2.4.1 Dados de leucemia

Foi considerado o conjunto de dados descritos no Exemplo 2.1. O tempo observado Y é relacionado ao relapso ou morte do paciente, em que, para cada indivíduo, o tempo

observado foi medido em dias, transformado em anos e, além disso, a proporção de indivíduos censurados é de 60.38%.

Foram ajustados os modelos de longa duração apresentados na Tabela 2.3 e para o modelo de longa duração binomial foram fixados valores para o parâmetro K no conjunto $\{0, 1, 2, \ldots, 40\}$. Também para o parâmetro τ do modelo de longa duração binomial negativa foi considerado que toma valores no conjunto $\{1, \ldots, 30\}$. Para a escolha do valor do parâmetro K no modelo de longa duração binomial foi observado que, de acordo com o critério AIC e BIC, quando $K = 30$ tem um melhor ajuste em relação a valores menores que 30. Foi observado que, para valores maiores que 30 no parâmetro K, a diferença entre os critérios AIC (BIC) é pequena e desta forma adotou-se $K = 30$ para o modelo de longa duração binomial. No caso do modelo de longa duração binomial negativa o melhor ajuste aos dados é quando $\tau = 1$, isto levando-se em conta o critério AIC e BIC. Por isso, conclui-se que o melhor ajuste é dado quando o número de causas competitivas segue a distribuição geométrica (MLDGeo).

Na Tabela 2.4, são apresentados os valores do máximo valor do logaritmo da verossimilhança, $\max \ell(.)$ e os valores dos critérios AIC e BIC considerando-se todas as covariáveis (modelo completo). Assim, o modelo que tem um melhor ajuste entre os modelos propostos, considerando-se os critérios AIC e BIC, é o modelo que considera a distribuição logarítmica como modelo para o número de riscos competitivos. Também observa-se que modelo de longa duração binomial ($K = 30$) e Poisson têm, de acordo com as estatísticas AIC e BIC ajustes muito próximos.

Tabela 2.4: Critérios de comparação de modelos para o conjunto de dados de leucemia.

Modelo	$\max \ell(.)$	AIC	BIC
MLDBer	-1750,14	3510,28	3537,66
MLDBi†	-1741,63	3493,25	3520,63
MLDPoi	-1741,36	3492,73	3520,10
MLDGeo	-1734,72	3479,43	3506,81
MLDLg	-1731,92	3473,92	3501,30

$^{\dagger} K = 30$

Para os modelos de longa duração estudados observou-se que as covariáveis são relacionadas com a fração de cura por meio do parâmetro de potência θ, assim, uma questão importante é saber quais fatores de risco têm influência na fração de cura. Neste cenário, para selecionar as covariáveis para os modelo de longa duração propostos serão adotados os critérios AIC e BIC. Neste contexto, é importante notar que o conjunto de dados de leucemia aguda tem 4 covariáveis, o que significa que existem $2^4 - 1 = 15$ modelos diferentes (combinação de covariáveis) para o modelo de longa duração de série de potências. Assim, para saber quais covariáveis têm influência na "cura"dos pacientes de leucemia, foi considerado o modelo de MLDLg que teve melhor ajuste aos dados.

A Tabela 2.5 apresenta os cinco melhores ajustes para o modelo MLDLg, assim pode ser observado que os critérios AIC e BIC indicam que as covariáveis ano de transplante de médula óssea (x_1) e idade do paciente (x_2) são fatores que têm influência na fração de cura no modelo MLDLg. É importante observar que as covariáveis x_1 e x_2 estão presentes na maioria dos modelos.

Tabela 2.5: Seleção de covariáveis para o conjunto de leucemia para o modelo MLDLg.

Covariaveis	AIC	Covariáveis	BIC
x_1, x_2	3473,92	x_1, x_2	3501,30
x_1, x_2 , x_3	3474,69	x_1	3506,98
x_1, x_2, x_4	3475,10	x_1, x_2 , x_3	3507,54
x_1, x_2, x_3, x_4	3475,77	x_1, x_2, x_4	3507,95
x_1, x_3	3484,82	x_1, x_3	3512,20

Na Tabela 2.6 são apresentadas as estimativas de máxima verossimilhança e os erro padrões para o modelo MLDLg.

Tabela 2.6: Estimativas de máxima verossimilhança dos parâmetros do modelo MLDLg e os erro padrões para o conjunto de dados de leucemia.

Parâmetro	Estimativa	Erro Padrão
α	0,922	0,022
λ	-0,518	0,060
β_0	0,937	0,151
β_1	-0,747	0,142
β_2	0,522	0,126

A Figura 2.4 mostra as estimativas de K-M da função de sobrevivência (linhas contínuas), assim como as estimativas do modelo MLDLg para a covariável idade (figura esquerda) e ano de transplante de médula óssea (figura direita). Considerando o grupo de pacientes que têm idade menor ou igual a 20, observa-se que a estimativa da sobrevivência, de acordo com o modelo MLDLg, é muito próxima da curva de K-M e de forma análoga a estimativa da função de sobrevivência para o grupo de pacientes com idades entre 20 e 40 anos do modelo MLDLg é próxima do curva K-M, pelo menos nos 10 anos de acompanhamento dos pacientes.

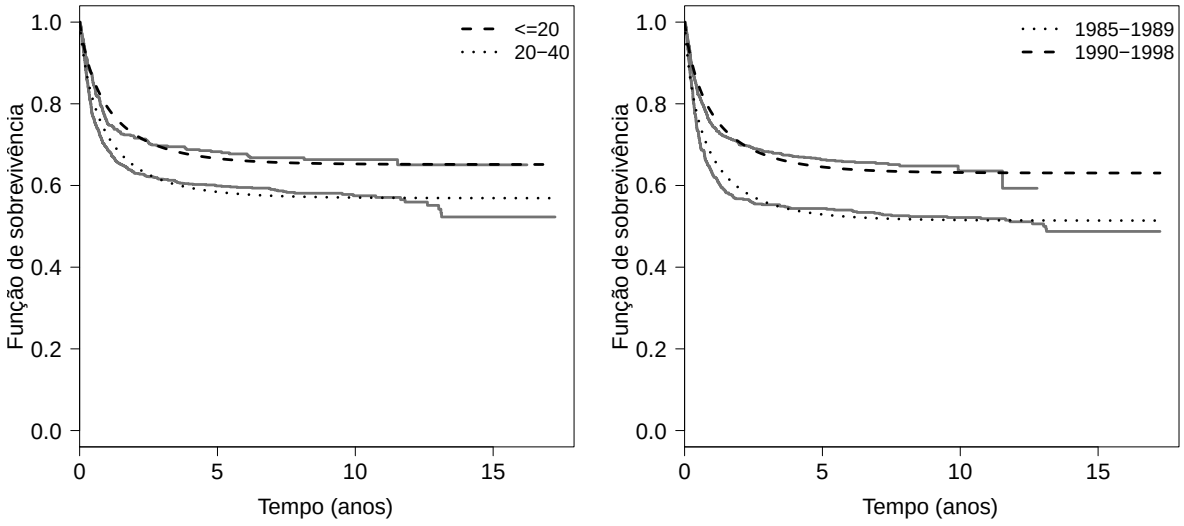


Figura 2.4: Estimativa de K-M da função de sobrevivência e estimativa da função de sobrevivência estratificado para as covariáveis idade (painel esquerdo) e ano de transplante (painel direito), de acordo com o modelo MLDLg para os dados de pacientes com leucemia.

O lado direito da Figura 2.4, exibe uma aproximação as curva de K-M para os diferentes anos de transplante, porém pode-se observar que o estrato de pacientes que tiveram o transplante entre os anos 1990-1998, tem um acompanhamento menor, embora o ajuste do modelo MLDLg consiga uma razoável aproximação a curva de K-M.

Apresenta-se na Tabela 2.7 estimativas da fração de cura (em media) para as covaráveis x_1 e x_2. Observa-se que a probabilidade de ser curado é maior quando o transplante da médula é feita nos anos 1990 até 1998. Além disso, a probabilidade de cura é maior em pacientes que receberam o transplante de médula óssea quando a idade de eles foi menor ou igual a 20 anos.

Tabela 2.7: Estimativa da fração de cura para o conjunto de dados de leucemia.

Modelo	Ano de transplante (x_1)		Idade (x_2)	
	$1985-1989$	$1990-1998$	≤ 20	$20-40$
MLDBer	0,508	0,648	0,666	0,575
MLDPoi	0,502	0,643	0,663	0,569
MLDGeo	0,503	0,636	0,656	0,566
MLDLg	0,512	0,629	0,649	0,566

2.4.2 Dados de melanoma

Para o conjunto de dados de melanoma apresentado no Exemplo 2.2, foram ajustados os modelos de longa duração apresentados na Tabela 2.3. Para o modelo MLDBi, foi considerado um conjunto de valores para o parâmetro K, assim $K \in \{1, 2, \ldots, 200\}$. No modelo MLDBn, foi assumido que o parâmetro τ toma valores no conjunto $\{1, \ldots, 20\}$.

De acordo com os critérios AIC e BIC, o melhor ajuste para o modelo MLDBi, considerando-se o conjunto de valores que toma o parâmetro K, é dado quando $K = 191$.

Para o modelo MLDBn, o melhor ajuste aos dados é quando $\tau = 1$ levando-se em conta os critérios AIC e BIC e, em seguida, o melhor ajuste é dado quando o número de causas competitivas segue a distribuição geométrica (MLDGeo).

A Tabela 2.8 apresenta os valores do máximo valor do logaritmo da verossimilhança, $\max \ell(.)$ e os valores dos critérios AIC e BIC considerando-se todas as covariáveis (modelo completo).

Neste cenário, o modelo de longa duração que tem um melhor ajuste entre os modelos propostos (Seção 2.2), de acordo com os critérios AIC e BIC é o modelo MLDLg. Observe-se que o modelo MLDBi ($K = 191$) e o MLDPoi tem ajuste próximos. Uma situação similar foi observada quando é considerado o conjunto de dados de leucemia.

Para selecionar as covariáveis que têm efeito na fração cura, foram considerados os critérios AIC e BIC. Considerando-se que o conjunto de dados de melanoma tem seis

Tabela 2.8: Critérios de comparação de modelos para o conjunto de dados de melanoma.

Modelo	$\ell(\cdot)$	AIC	BIC
MLDBer	-513,79	1045,58	1081,88
MLDBi†	-510,06	1038,13	1074,43
MLDPoi	-510,05	1038,10	1074,39
MLDGeo	-506,88	1031,76	1068,06
MLDLg	-505,28	1028,57	1064,86

$^{\dagger}K = 191$

covariáveis, então tem-se $2^6 - 1 = 63$ modelos diferentes (combinação de covariáveis) para o modelo de longa duração de série de potências. Os modelos MLDLg e MLDGeo são os modelos de longa duração que têm um melhor ajuste para o conjunto de dados, e levando-se em conta isto, será adotado o modelo MLDLg para selecionar que covariáveis têm influência na fração de cura.

Na Tabela 2.9 são apresentados os cinco melhores ajustes para o modelo MLDLg. Assim, observa-se que o melhor modelo escolhido, considerando-se o critério AIC, é o modelo dado pelas covariáveis idade (x_2) e categoria do nódulo (x_3) e, seguindo o critério BIC, tem-se que é o modelo composto só com a covariável x_3. Nota-se também que a variável x_3 está presente nos 5 melhores modelos para ambos critérios.

Tabela 2.9: Seleção de covariáveis para o conjunto de melanoma para o modelo MLDLg.

Covariaveis	AIC	Covariaveis	BIC
x_2, x_3	1023,40	x_3	1042,05
x_2, x_3, x_6	1023,58	x_2, x_3	1043,56
x_2, x_3, x_5	1024,70	x_3, x_6	1046,21
x_1, x_2, x_3	1024,96	x_3, x_5	1046,96
x_2, x_3, x_5, x_6	1025,05	x_1, x_3	1047,30

Na Tabela 2.10 são apresentadas as estimativas de máxima verossimilhança e os erro padrões para o modelo MLDLg considerando as covariáveis x_2 e x_3.

Tabela 2.10: Estimativas de máxima verossimilhança dos parâmetros do modelo MLDLg e os erro padrões para o conjunto de dados de melanoma

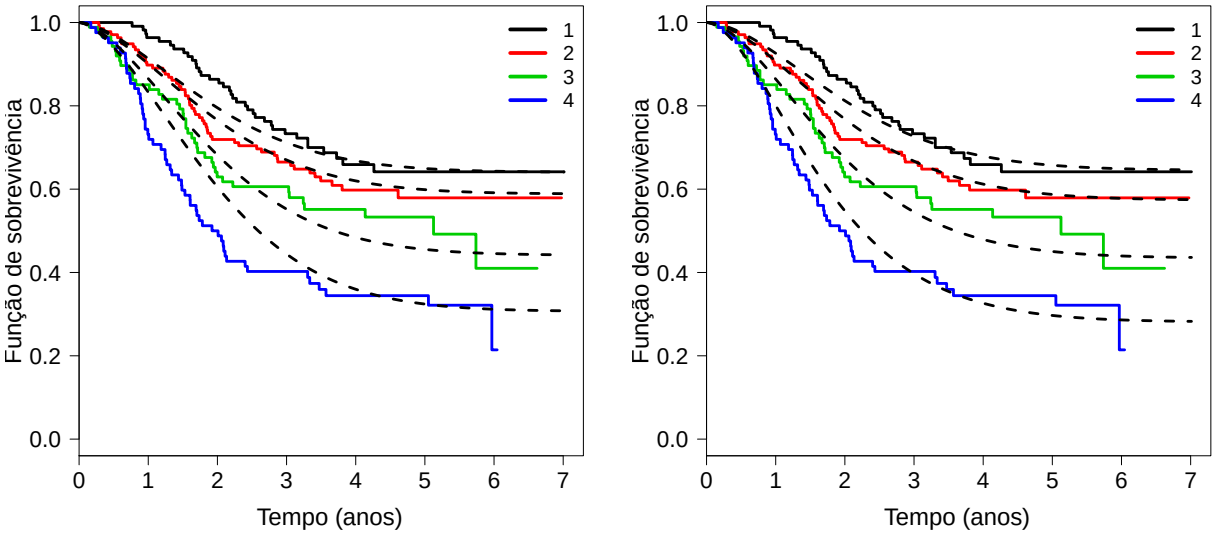
Parâmetro	Estimativa	Erro padrão
α	2,089	0,135
λ	-2,489	0,206
β_0	-1,526	0,651
β_2	0,023	0,011
β_3	0,800	0,144

Apresenta-se na Tabela 2.11 estimativas da fração de cura (em media) para a covarável x_3. Observa-se que a probabilidade de cura é menor se o indivíduo esta no estado mais severo do câncer isto é ele pertence a categoria 4. Porém, se o paciente esta no estagio inicial da doença tem-se que a probabilidade de ser curado é maior.

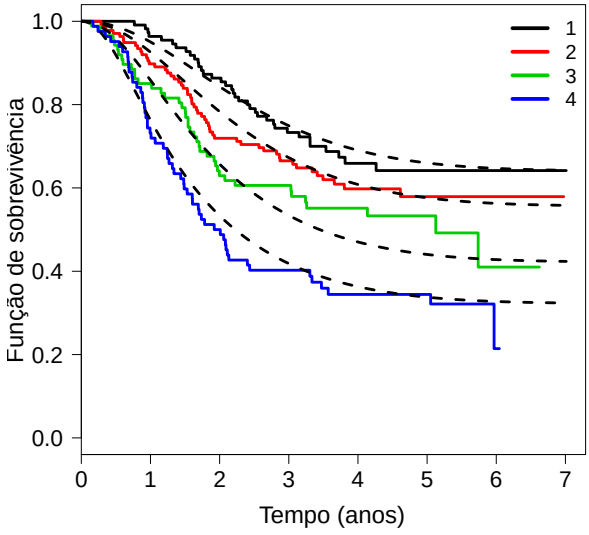
Tabela 2.11: Estimativas da fração de cura para o conjunto de dados de melanoma considerando a covariável x_3.

Modelo	Categoria do nodulo (x_3)			
	1	2	3	4
MLDBer	0,660	0,565	0,448	0,322
MLDPoi	0,659	0,564	0,440	0,293
MLDGeo	0,655	0,552	0,427	0,299
MLDLg	0,647	0,536	0,424	0,329

A Figura 2.5 mostra as estimativas de K-M da função de sobrevivência (linhas contínuas), assim como as estimativas dos modelos MLDBer, MLDPoi e MLDLg (linhas pontilhadas) para a covariável categoria do nódulo (x_3)(1,2,3,4). Observe-se que o ajuste do modelo MLDBer apresentado na Figura 2.5(a) não tem um ajuste satisfatório. Em relação aos modelos MLDBer e MLDPoi (ver Figura 2.5(b)) nota-se que o modelo MLDLg da Figura 2.5(c) fornece uma melhor aproximação das curvas de K-M pelo menos nos primeiros anos de estudo.



(a) (b)



(c)

Figura 2.5: Estimativa de K-M e paramétricas da função de sobrevivência de acordo com a covariável categoria do nodulo (x_3): (a) MLDBer, (b) MLDPoi e (c) MLDLg - Exemplo 2.2

2.5 Comentários finais

Neste capítulo, foi apresentado o modelo de série de potências com fração de cura para modelar dados de sobrevivência de longa duração de uma perspectiva frequentista considerando covariáveis. Neste sentido, o modelo MLDDSP é flexível e inclui outros

modelos de longa duração. Na aplicação do modelo MLDDSP para o conjunto de dados de leucemia e de melanoma, observa-se que o modelo MLDLg se ajusta melhor e, além disso, também foram considerados os critérios AIC e BIC para a seleção de variáveis preditoras na fração de cura. Neste sentido, as covariáveis selecionadas de acordo com os critérios AIC e BIC são similares, como pode ser visto na Tabela 2.5, para os pacientes com leucemia e na Tabela 2.9, para os pacientes com melanoma.

Uma extensão para os modelos de longa duração estudados neste capítulo pode ser considerada por incluir o termo de fragilidade na função de risco que, por exemplo em Gonzales *et al.* (2013), foi considerado o modelo de mistura padrão com fragilidade gama.

Foi considerado a linearidade das variáveis preditoras na função de ligação logito e logarítmica para os modelos MLDBi, MLDBn, MLDLg e MLDPoi respectivamente. Porém essa suposição, isto é, a linearidade das covariáveis em algumas situações pode ser questionável como pode ser visto em Friedman (1991) e Holmes & Mallick (2003). Com objetivo de mostrar uma abordagem capaz de capturar não linearidade das covariáveis será apresentado no Capítulo 3 uma metodologia baseada em partição.

Capítulo 3

Modelo de partição bayesiana

Modelagem de dados baseados em partição não é uma ideia nova, existem várias áreas da ciência em que esse tipo de abordagem é aplicada, tais como: epidemiologia, genética, geoestatística, finanças, entre outras. Em estatística espacial, por exemplo, um dos problemas relevantes é a estimação da incidência ou risco de uma certa doença em uma região de interesse. Nesse sentido, a ideia de fazer uma partição dessa região e analisar cada sub-região é geralmente utilizada em estatística espacial.

Assim, seja $\boldsymbol{\mathcal{X}}$ um domínio de interesse e seja uma família de subconjuntos $R_1, \ldots, R_M$ de $\boldsymbol{\mathcal{X}}$ ($R_m \neq \emptyset$), esta familia de subconjuntos define uma partição em $\boldsymbol{\mathcal{X}}$ se satisfaz

$$\bigcup_{m=1}^{M} R_m = \boldsymbol{\mathcal{X}} \quad \text{e} \quad R_{m'} \cap R_m = \emptyset \quad \text{se} \quad m \neq m'.$$

Na literatura estatística existem vários modelos baseados na ideia de partição, por exemplo, Barry & Hartigan (1992) propõem um modelo para identificar pontos de mudança considerando o modelo de partição produto proposto por Hartigan (1990).

A ideia de partição é também utilizada em regressão não paramétrica, por exemplo, o modelo de regressão por árvore e o modelo de regressão adaptativa multivariável por *splines* (do inglês, multivariate adaptive regression splines) (Friedman, 1991) fazem a partição no espaço das covariáveis. Nesse sentido, modelos de regressão por árvore utilizam uma árvore binária para dividir o espaço preditor e são amplamente utilizados em aprendizado de máquina (do inglês, machine learnnig) e mineração de dados (do inglês, data mining). Neste contexto, os modelos apresentados neste trabalho também consideram a partição no espaço das covariáveis em que é utilizado a modelagem de partição bayesiana. A vantagem desta proposta é capturar a não linearidade das covariáveis e além disso selecionar as covariáveis que tem influencia na variável resposta.

A seguir é considerado a definição de modelo de partição dada em Denison *et al.* (2002b)

Definição 3.1. Um modelo de partição é composto por um número de regiões disjuntas $R_1, \ldots, R_M$ cuja união é o domínio de interesse $\boldsymbol{\mathcal{X}}$, tal que $R_m \cap R_{m'} = \emptyset$, para $m \neq m'$ e $\bigcup_{m=1}^{M} R_m = \boldsymbol{\mathcal{X}}$. As respostas em cada região, dado o vetor de parâmetros relativo a cada região $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M)$, são permutáveis e provêm de uma mesma classe de distribuição f.

Muitos modelos de partição propostos na literatura em geral assumem a dependência dos parâmetros entre regiões próximas, como pode ser visto em Heikkinen & Arjas (1998), Heikkinen & Arjas (1999) entre outros. Porém, considerar dependência, conduz a alguns problemas e, neste sentido, Holmes *et al.* (1999, 2005) descrevem os problemas que levam a assumir a dependência na modelagem por partição, que são:

1. Considerando-se dependência dos parâmetros entre regiões próximas, tem-se como consequência, a dependência entre regiões próximas em $\boldsymbol{\mathcal{X}}$, portanto a dependência precisa ser especificada. Isto torna o modelo mais complexo, e, portanto, com mais parâmetros.

2. A função de verossimilhança marginal para a estrutura de partição, não é analiticamente tratável quando a dependência é adotada.

Além disso, considerando-se uma abordagem bayesiana em modelos que consideram a partição do espaço preditor, por exemplo, o modelo de classificação e regressão por árvore (CART), tem-se dificuldade na simulação MCMC (Monte Carlo via cadeias de Markov) para obter amostras distribuição *a posteriori* do modelo, isto é, o amostrador MCMC fica preso em uma moda local. Outras desvantagens da abordagem bayesiana do modelo CART foram discutidas em Chipman *et al.* (1998) e Denison *et al.* (1998).

Nesse cenário, Holmes *et al.* (1999, 2005) propõem um modelo de partição que assume independência entre os parâmetros de regiões próximas e não dividem o espaço preditor usando uma estrutura hierárquica como é feito em modelos baseados em árvores binárias (e.g., CART). Assim, para fazer a partição do espaço preditor $\boldsymbol{\mathcal{X}}$, Holmes *et al.* (1999, 2005) usam a tesselação de Voronoi. Seguindo essa ideia, uma tesselação de um conjunto $\boldsymbol{\mathcal{X}} \subseteq \mathbb{R}^p$ é uma coleção de regiões chamadas células, de forma que as células são disjuntas entre si e a união delas é conjunto $\boldsymbol{\mathcal{X}}$. A construção da tesselação é feita quando a forma geométrica

de cada célula é um polígono (ou uma região aberta) em que, não necessariamente, todas as células têm o mesmo número de lados. Uma boa referência de estudo e aplicações de diferentes tipos de tesselações pode ser vista em Stoyan *et al.* (1995) e Okabe *et al.* (2000).

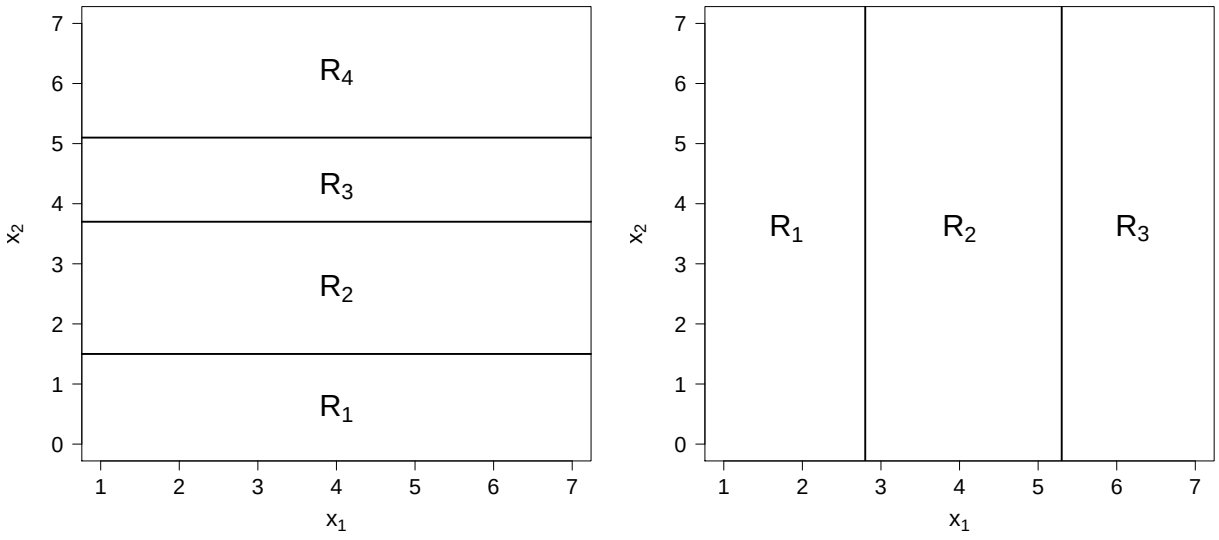
Levando-se em conta a independência dos parâmetros entre as regiões e fazendo-se uma partição do espaço preditor, $\boldsymbol{\mathcal{X}}$, sem considerar uma estrutura hierárquica, Holmes *et al.* (1999, 2005) propõem o modelo de partição bayesiana (MPB). Assim como o modelo CART, o modelo de partição bayesiana inicialmente foi proposto para problemas de regressão e classificação, porém algumas extensões do modelo MPB foram aplicadas para mapeamento de doenças (Denison & Holmes, 2001).

É importante observar que Holmes *et al.* (1999, 2005) considerou variáveis preditoras de natureza contínua para desenvolver o MPB. Porém, neste trabalho além de considerar covariáveis contínuas, também foram consideradas covariáveis categóricas, aplicados a modelos de sobrevivência com fração de cura e portanto uma nova estratégia computacional é proposta.

3.1 Modelo de partição bayesiana com hiperplanos

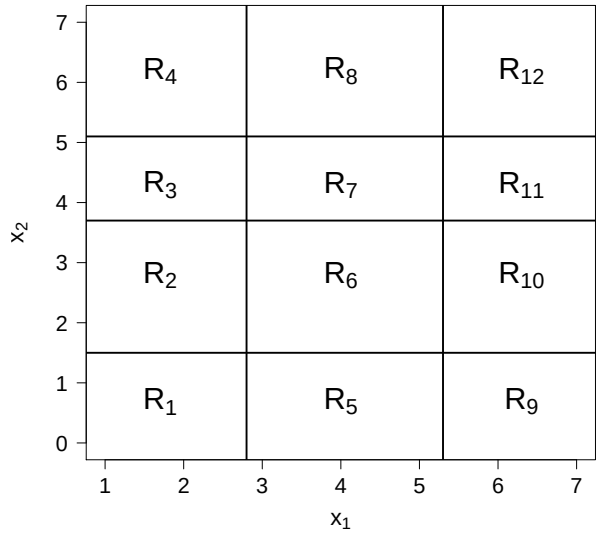
Neste trabalho foi adotada a tesselação por hiperplanos ortogonais aos eixos. Nesse sentido, para visualizar melhor essa tesselação, foi apresentada na Figura 3.1 uma tesselação por hiperplanos ortogonais em que a região de interesse é dada por $\boldsymbol{\mathcal{X}} = [0, 7] \times [0, 7]$. Para a construção da tesselação, observe-se que, na Figura 3.1(a), tem-se três pontos no eixo x_2 e, em seguida, eles determinam hiperplanos (em 1 dimensão são retas) paralelos ao eixo x_1. Na Figura 3.1(b) tem-se dois pontos no eixo x_1 que determinam 3 hiperplanos paralelos ao eixo x_2 e, desta forma, esses hiperplanos determinam uma partição de $\boldsymbol{\mathcal{X}}$, como pode ser visto na Figura 3.1(c).

Denota-se $\boldsymbol{\mathcal{T}}$ como sendo uma tesselação por hiperplanos ortogonais que determina M regiões disjuntas em $\boldsymbol{\mathcal{X}}$, em que as regiões são subconjuntos do espaço preditor e são denotadas por R_m $m = 1, \ldots, M$. Seja Y uma variável resposta com vetor $(p \times 1)$ de variáveis preditoras, então o modelo de partição bayesiana atribui um modelo paramétrico para Y em cada região do $\boldsymbol{\mathcal{X}}$ sendo que $\boldsymbol{x}$ situa-se na m ésima região, R_m, e logo se assume que Y segue um modelo paramétrico, $Y \sim f(y|\boldsymbol{\theta}_m)$, indexado pelo parâmetro $\boldsymbol{\theta}_m$ (parâmetro local). Assim o modelo MPB é determinado por duas componentes: a estrutura de tesselação $\boldsymbol{\mathcal{T}}$ que divide em M regiões o espaço preditor e o vetor de parâmetros do



(a)

(b)



(c)

Figura 3.1: (a) Retas paralelas ao eixo x_1, (b) Retas paralelas ao eixo x_2 e (c) Retas ortogonais aos eixos x_1 e x_2.

modelo assumido em cada região $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M)^\top$.

Para indicar que uma variável resposta Y com vetor de covariáveis $\boldsymbol{x}$ são associados à m-ésima região R_m, adotou-se a notação Y_{mj} para a variável resposta e $\boldsymbol{x}_{mj}$ para o vetor de covariáveis respectivamente, $m = 1, \ldots, M$, $j = 1, \ldots, n_m$ em que n_m representa o número de pontos em R_m e, considerando-se uma amostra de tamanho n, então $\sum_{m=1}^{M} n_m = n$.

Denota-se o conjunto de variáveis respostas e suas respectivas covariáveis que pertencem à região R_m por $\boldsymbol{Y}_m$ e $\boldsymbol{X}_m$ em que

$$\boldsymbol{Y}_m = (y_{m1}, \ldots, y_{mn_m}) \quad \text{e} \quad \boldsymbol{X}_m = (\boldsymbol{x}_{m1}, \ldots, \boldsymbol{x}_{mn_m}), \quad m = 1, \ldots, M. \tag{3.1}$$

Dada a tesselação $\boldsymbol{\mathcal{T}}$ e os parâmetros locais em cada região, a função de verossimilhança para uma amostra de tamanho n é dada por

$$L(\boldsymbol{\mathcal{T}}, \boldsymbol{\theta}|\mathcal{D}) = \prod_{m=1}^{M} \prod_{j=1}^{n_m} f(y_{mj}|\theta_m), \tag{3.2}$$

em que $\mathcal{D} = \{y_i, \boldsymbol{x}_i\}_1^n$.

Devido a natureza da variável resposta Y, pode-se considerar diferentes modelos paramétricos para $Y \sim f(y|\theta_m)$ em cada região do espaço preditor. Geralmente são considerados um modelo de regressão linear ou no caso de problemas de classificação é usada a distribuição multinomial (Chipman *et al.*, 1998; Holmes *et al.*, 2005).

3.1.1 Especificação *a priori* para o modelo de partição bayesiana

Uma vez que o modelo de partição bayesiana é determinado por $(\boldsymbol{\mathcal{T}}, \boldsymbol{\theta})$ e considerando-se uma abordagem bayesiana para a análise dos dados, é necessário especificar a distribuição *a priori* para $p(\boldsymbol{\mathcal{T}}, \boldsymbol{\theta})$. Primeiramente, especifica-se a distribuição *a priori* para a tesselação $\boldsymbol{\mathcal{T}}$ e, em seguida, especifica-se a distribuição *a priori* para os parâmetros locais $p(\boldsymbol{\theta}|\boldsymbol{\mathcal{T}})$.

A estrutura da tesselação $\boldsymbol{\mathcal{T}}$ é composta pelos hiperplanos $\boldsymbol{h}$ paralelos aos eixos e o número M de regiões em $\boldsymbol{\mathcal{X}}$, assim $\boldsymbol{\mathcal{T}} = \{\boldsymbol{h}, M\}$. Observa-se que $\mathcal{I}$ é o conjunto dos índices das variáveis preditoras $\mathcal{I} = \{1, \ldots, p\}$ e $\mathcal{I}_{\boldsymbol{\mathcal{T}}}$ é o conjunto de índices das covariáveis presentes na tesselação $\boldsymbol{\mathcal{T}}$.

Supondo-se que existam p^* covariáveis presentes em $\mathcal{I}_{\boldsymbol{\mathcal{T}}}$, tem-se que, para cada variável preditora em $\mathcal{I}_{\boldsymbol{\mathcal{T}}}$ existe, pelo menos, um hiperplano ortogonal a essa variável preditora. Neste sentido, a tesselação faz uma partição no conjunto composto pelas p^* covariáveis das n unidades amostrais, assim esse conjunto é geralmente considerada como o envelope convexo formada pelas p^* covariáveis.

Seja $\boldsymbol{h}_{r^*}$ que denota o vetor de pontos de corte na covariável r^*, $r^* \in \mathcal{I}_{\boldsymbol{\mathcal{T}}}$ em seguida, é importante notar que a escolha dos pontos de corte em cada covariável é feita aleatoriamente e considerando-se que os pontos são uniformemente distribuídos sob conjunto de valores observados para essa variável preditora r^*, a distribuição *a priori* para o vetor de

hiperplanos $\boldsymbol{h}$ é dada por

$$p(\boldsymbol{h}) = \prod_{r^*=1}^{p^*} \mathrm{U}(\boldsymbol{h}_{r^*}|\boldsymbol{x}_{r^*}). \tag{3.3}$$

A finalidade de considerar uma distribuição *a priori* uniforme sob os valores para cada uma das p^* covariáveis para os pontos de corte é evitar regiões vazias. No cenário que os dados apresentam só uma covariável, o modelo de partição bayesiana pode ser relacionado com os modelos de ponto de mudança (Denison *et al.*, 2002b). Em esse sentido, os conjuntos de dados considerados neste trabalho tem mais de duas covariáveis e identificar pontos de mudança, por exemplo, em covariáveis contínuas torna-se uma tarefa difícil. Porém, se a covariável for qualitativa com mais de dois categorias pode-se conhecer *a priori* todos os possíveis agrupamentos dessa covariável. Em seguida, é possível identificar que agrupamentos são os mais plausíveis para o modelo proposto. Assim na Seção 3.1.3 apresenta-se a maneira de como uma covariável categórica vai ser particionada.

Para o número de regiões, M, no espaço preditor pode-se considerar diferentes distribuições *a priori* . Neste trabalho, foi adotada como distribuição *a priori* para M a distribuição geométrica com média $1/\psi$

$$p(M) = \psi(1-\psi)^{M-1}, \quad M = 1, 2, \ldots, \tag{3.4}$$

e a distribuição *a priori* para a estrutura da tesselação $\boldsymbol{\mathcal{T}}$ é dada por

$$p(\boldsymbol{\mathcal{T}}) = p(\boldsymbol{h})p(M). \tag{3.5}$$

Na prática o número de regiões em $\mathcal{X}$ não pode ser maior que o tamanho da amostra ($M < n$). A distribuição *a priori* para $\boldsymbol{\mathcal{T}}$, definida em (3.5), será adotada para todas as aplicações do presente trabalho, independentemente da distribuição assumida para a variável resposta Y.

Neste contexto, especificar uma distribuição *a priori* para o vetor de parâmetros associado com a distribuição paramétrica, $f(y|\theta)$, em cada região, depende basicamente do modelo assumido. Porém é possível considerar algumas especificações, como por exemplo, adotar modelos em que seja possível uma simplificação analítica para que o esforço computacional seja menor e desta forma se possam obter amostras da distribuição *a posteriori* para o modelo de partição bayesiana. Nesse sentido, considerar distribuições *a priori* conjugadas para o vetor de parâmetros em cada região de $\boldsymbol{\mathcal{X}}$ é importante para que a simplificação analítica seja possível. No entanto, nem todos os modelos

paramétricos assumidos para Y têm uma distribuição conjugada, assim outras ferramentas computacionais podem ser empregadas, como integração numérica (Kim *et al.*, 2005).

Seja $p(\theta_m)$ a distribuição *a priori* para os parâmetros locais em R_m. Devido à tesselação $\boldsymbol{\mathcal{T}}$ com M regiões em $\boldsymbol{\mathcal{X}}$, tem-se a que a distribuição *a priori* para $\boldsymbol{\theta}$ é dada por

$$p(\boldsymbol{\theta}|\boldsymbol{\mathcal{T}}) = \prod_{m=1}^{M} p(\theta_m), \tag{3.6}$$

em que é importante observar a independência dos parâmetros entre as regiões.

3.1.2 Análise *a posteriori*

Considerando-se as distribuições *a priori* para a tesselação $\boldsymbol{\mathcal{T}}$ e o vetor de parâmetros locais $\boldsymbol{\theta}$ e a função de verossimilhança, tem-se que a distribuição *a posteriori* conjunta para $\{\boldsymbol{\mathcal{T}}, \boldsymbol{\theta}\}$ é dada por

$$p(\boldsymbol{\mathcal{T}}, \boldsymbol{\theta}|\mathcal{D}) \propto L(\boldsymbol{\mathcal{T}}, \boldsymbol{\theta}|\mathcal{D}) p(\boldsymbol{\theta}, \boldsymbol{\mathcal{T}}). \tag{3.7}$$

O tratamento analítico da distribuição *a posteriori* é difícil, por isso, para obter amostras da distribuição *a posteriori* , foi usado um amostrador MCMC. Além disso, observe-se que a distribuição *a posteriori* conjunta dada em (3.7) pode ser fatorizada sempre na forma seguinte

$$p(\boldsymbol{\mathcal{T}}, \boldsymbol{\theta}|\mathcal{D}) = p(\boldsymbol{\mathcal{T}}|\mathcal{D}) p(\boldsymbol{\theta}|\boldsymbol{\mathcal{T}}, \mathcal{D}), \tag{3.8}$$

e, em seguida, observa-se que a distribuição *a posteriori* conjunta para o modelo de partição é o produto da probabilidade *a posteriori* da tesselação e a distribuição condicional dos parâmetros do modelo adotado (em cada região).

Para o cálculo da distribuição *a posteriori* de $p(\boldsymbol{\mathcal{T}}|\mathcal{D})$ é necessário obter a verossimilhança marginal, para qualquer estrutura de tesselação $\boldsymbol{\mathcal{T}}$. Nesse sentido, é importante notar que, ter uma forma fechada da verossimilhança marginal para $\boldsymbol{\mathcal{T}}$ depende basicamente da escolha da distribuição *a priori* para os parâmetros locais em cada região do espaço preditor,$\boldsymbol{\mathcal{X}}$. Assim sendo que foram atribuídas distribuições *a priori* conjugadas para os parâmetros locais θ_m, e desta forma, obteve-se a verossimilhança marginal

$$L(\boldsymbol{\mathcal{T}}|\mathcal{D}) = \int L(\boldsymbol{\mathcal{T}}, \boldsymbol{\theta}|\mathcal{D}) p(\boldsymbol{\theta}|\boldsymbol{\mathcal{T}}) \mathrm{d}\boldsymbol{\theta}, \tag{3.9}$$

portanto a distribuição *a posteriori* para $\boldsymbol{\mathcal{T}}$ é dada por

$$p(\boldsymbol{\mathcal{T}}|\mathcal{D}) \propto L(\boldsymbol{\mathcal{T}}|\mathcal{D}) p(\boldsymbol{\mathcal{T}}), \tag{3.10}$$

em que para explorar a distribuição a posterior de $\boldsymbol{\mathcal{T}}$, usamos um algoritmo MCMC. Não obstante, pode-se ter o cenário em que não exista uma distribuição *a priori* conjugada para o parâmetro local. Nesse caso, para o calculo da verossimilhança marginal, $L(\boldsymbol{\mathcal{T}}|\mathcal{D})$, pode-se utilizar, por exemplo, integração numérica (Kim *et al.*, 2005) porém o custo computacional é maior.

Dado que os parâmetros locais são integrados em (3.9) tem-se que o número dos modelos a ser explorados na distribuição *a posteriori* se reduz e além disso o amostrador MCMC tem uma melhor performance como pode ser visto em Han & Carlin (2001).

Neste contexto, para obter amostras da distribuição *a posteriori* de $\boldsymbol{\mathcal{T}}$, Holmes *et al.* (1999, 2005) propõem um amostrador MCMC com saltos reversíveis baseado no método Monte Carlo via cadeias de Markov com saltos reversíveis (RJMCMC) (Green, 1995). Porém um amostrador MCMC diferente de Holmes *et al.* (2005) é proposto neste trabalho pelo fato que a tesselação adotada é baseada em hiperplanos ortogonais e será apresentada na Seção 3.1.3.

3.1.3 Estratégia computacional

A estratégia computacional para covariáveis quantitativas e dicotômicas foi considerada por Hoggart & Griffin (2001). Em geral, covariáveis qualitativas não são, necessariamente dicotômicas, razão pela qual, neste trabalho foi modificada a estratégia computacional proposta por Hoggart & Griffin (2001) por considerar covariáveis qualitativas com mais de duas categorias.

Supondo que X_C seja uma variável preditora qualitativa com g categorias, $X_C \in \{1, 2, \ldots, g\}$ e denota-se por ρ sendo uma partição de X_C e seja M_ρ o número de subconjuntos (grupos) de X_C de acordo com a partição ρ em que ρ é desconhecida. Foi adotada a distribuição uniforme sobre $\{1, \ldots, n_\rho\}$ como a distribuição *a priori* para ρ, em que n_ρ é o número de diferentes partições de X_C.

Na Tabela 3.1 são apresentados o número de grupos em X_C e o número total de partições considerando que $g = 4$, $X_C \in \{1, 2, 3, 4\}$. Assim no caso que $M_\rho = 2$ tem-se que existem sete possíveis partições para X_C isto se não for considerando a ordem da covariável assim uma partição da covariável X_C pode ser por exemplo $\rho = \{\{1, 3\}, \{2, 4\}\}$.

Porém se X_C é uma variável categorica ordinal então tem-se que assumir a ordem nos agrupamentos. Assim se $M_\rho = 2$ tem-se três possíveis diferentes partições. Portanto, no caso em que X_C é uma variável qualitativa ordinal o número de diferentes partições ρ

com M_ρ subconjuntos é menor em relação quando X_C é uma variável qualitativa nominal (McCullagh & Yang, 2008).

Tabela 3.1: Numero de subconjuntos e de partições se de X_C se $g = 4$

M_ρ	ρ (sem ordem)	ρ (com ordem)
1	1	1
2	7	3
3	6	3
4	1	1
n_ρ	15	8

Na Tabela 3.2 são apresentadas as partições de X_C, em que se leva em consideração a ordem das categorias. Observa-se que o número de agrupamentos com dois subconjuntos para X_C é três ($M_\rho = 3$).

Tabela 3.2: Número de partições de X_C (ordem).

M_ρ	Grupos
1	{1,2,3,4}
2	{1},{2,3,4}
	{1,2},{3,4}
	{1,2,3},{4}
3	{1},{2},{3,4}
	{1},{2,3},{4}
	{1,2},{3},{4}
4	{1},{2},{3},{4}

Uma vez que $\mathcal{I}$ representa o conjunto dos índices das covariáveis e $\mathcal{I}_{\boldsymbol{\mathcal{T}}}$ é o conjunto de índices das covariáveis presentes na tesselação $\boldsymbol{\mathcal{T}}$, o algoritmo proposto começa ($\mathcal{I}_{\boldsymbol{\mathcal{T}}} = \emptyset$) por escolher, aleatoriamente uma variável preditora e, em seguida, seleciona um ponto de corte dessa variável. Em cada iteração do algoritmo, e levando em conta que $1 < M < n$, escolhem-se os três primeiros movimentos. Os dois primeiros movimentos do algoritmo estão relacionados com a seleção da covariável. Os três últimos movimentos envolvem variáveis categóricas. De forma geral tem-se o seguinte algoritmo

- **Adição:** um novo hiperplano é adicionado à tesselação $\boldsymbol{\mathcal{T}}$ por escolher um novo ponto de corte de uma variável, sendo que o índice da variável está em $\mathcal{I}$. O ponto de corte é selecionado da distribuição empírica da variável escolhida.

- **Eliminação:** um hiperplano pode ser eliminado por escolher ao acaso uma variável preditora, r^*, presente na tesselação $r^* \in \mathcal{I}_{\boldsymbol{\mathcal{T}}}$.

- **Movimento:** um hiperplano pode ser mudado por selecionar outro ponto de corte da distribuição empírica da covariável escolhida em $\mathcal{I}_{\boldsymbol{\mathcal{T}}}$.

- **Combinação:** o número de grupos na covariável X_C decresce, por juntar dois grupos.

- **Divisão:** o número de grupos na covariável X_C cresce por dividir um grupo em dois novos subconjuntos.

- **Alteração:** a partição ρ de X_C é alterada embora o número M_ρ de grupos em ρ permaneça igual.

A tesselação proposta $\boldsymbol{\mathcal{T}}'$ é aceita com probabilidade

$$A(\boldsymbol{\mathcal{T}}', \boldsymbol{\mathcal{T}}) = \min\left\{1, \frac{L(\boldsymbol{\mathcal{T}}'|\mathcal{D})p(\boldsymbol{\mathcal{T}}')}{L(\boldsymbol{\mathcal{T}}|\mathcal{D})p(\boldsymbol{\mathcal{T}})}\right\}. \tag{3.11}$$

Note-se que a fração em (3.11) é o fator de Bayes em favor da novo modelo proposto, isto é, uma nova estrutura de partição.

É importante ressaltar que o algoritmo proposto anteriormente é um caso especial do método Monte Carlo via cadeias de Markov com saltos reversíveis (RJMCMC), isto pelo fato que foi assumido que foi possível calcular a verossimilhança marginal para $\boldsymbol{\mathcal{T}}$ (Green, 2003). Além disso, o Jacobiano da transformação requerido no algoritmo RJMCMC é identicamente igual a 1, isto pelo fato os pontos de corte (hiperplanos) são retirados de sua distribuição *a priori*, isto é, a distribuição empírica para cada covariável. Portanto, o amostrador MCMC proposto para explorar a distribuição *a posteriori* de $\boldsymbol{\mathcal{T}}$ é similar ao algoritmo Metropolis-Hastings. Nesse sentido, Denison *et al.* (2002b) discute as similaridades entre os algoritmos MCMC para simular a distribuição *a posteriori* dos modelos baseados em árvores binárias e o modelo de partição bayesiana proposto por Holmes *et al.* (2005). Uma diferença entre esses modelos esta na forma de fazer a partição do espaço preditor assim o modelo de partição bayesiana utiliza uma tesselação para dividir o espaço preditor e não faz uso de uma estrutura hierárquica como os modelos

que consideram uma árvore binaria para dividir o espaço preditor, por exemplo o modelo CART.

No algoritmo MCMC proposto tem-se que os três primeiros movimentos foram considerados por Hoggart & Griffin (2001). Neste trabalho adicionou-se os três últimos movimentos que estão ligados com às variáveis qualitativas. Para uma melhor compreensão dos três últimos movimentos para variáveis qualitativas, supõe-se que X_C tem 4 categorias, $X_C \in \{1, 2, 3, 4\}$. Seja ρ uma partição de X_C com três grupos, $M_\rho = 3$, $\rho = \{\{1, 3\}, \{2\}, \{4\}\}$. Observe-se que, neste caso em que $g = 4$ o número de grupos para uma partição ρ pode variar entre 1 e 4 subconjuntos.

Considerando-se o movimento **Combinação**, dois subconjuntos na partição ρ são unidos, a escolha dos grupos é aleatória e assim tem-se $\binom{M_\rho}{2}$ possibilidades para escolher dois grupos. Por exemplo, juntando-se os grupos $\{2\}$ e $\{4\}$, tem-se um novo grupo, por tanto isto leva a uma nova partição $\rho' = \{\{1, 3\}, \{2, 4\}\}$ com $M_{\rho'} = 2$.

Tendo-se o movimento **Divisão** no algoritmo, divide-se um grupo em dois subconjuntos. A escolha é feita aleatoriamente e restrita a subconjuntos com cardinalidade maior que 1. Assim, seja $\widetilde{M}_\rho$ o número de subconjuntos com mais que uma categoria na partição ρ de X_C. Na presente partição ρ' observa-se que $\widetilde{M}_{\rho'} = 2$ e supondo que foi escolhido $\{1, 3\}$ para fazer a divisão, tem-se que a nova partição é dada por $\rho = \{\{1\}, \{3\}, \{2, 4\}\}$ obtendo-se agora $M_\rho = 3$ grupos.

Para o movimento **Alteração**, supondo-se que a partição de X_C é dada por ρ, altera-se a configuração de ρ embora o número de grupos M_ρ não seja alterado. Na última partição tem-se que o número de grupos é $M_\rho = 3$, então uma configuração diferente para X_C pode ser, por exemplo, $\{\{1, 4\}, \{2\}, \{3\}\}$.

Na Tabela 3.1, observa-se que, se $M_\rho = 3$, tem-se 6 partições diferentes para X_C no caso que essa covariável seja uma variável qualitativa nominal.

Se a covariável qualitativa tem mais de 4 categorias o presente algoritmo pode ser aplicado, embora o custo computacional seja maior. No caso de assumir somente partições ordenadas, o número de partições diferentes diminui (Giudici *et al.*, 2000).

3.2 Alguns exemplos

Nesta seção, serão apresentados exemplos em que o modelo MPB é aplicado para alguns modelos considerando-se dados censurados e não censurados.

Exemplo 3.1. O modelo de regressão apresentado aqui é similar ao modelo de regressão por árvore proposto por Chipman *et al.* (1998) e Denison *et al.* (1998) considerando uma abordagem bayesiana. Não obstante, é utilizado a tesselação por hiperplanos ortogonais ao vez de uma árvore binária para dividir o espaço preditor.

Supondo que a tesselação $\boldsymbol{\mathcal{T}}$ particiona o espaço preditor em M subconjuntos, para fazer regressão adotamos que a variável Y resposta segue uma distribuição normal em cada região R_m, e tem-se

$$Y_{mj} \sim \mathrm{N}(\mu_m, \sigma^2), \quad m = 1, \ldots, M \quad j = 1, \ldots, n_m, \tag{3.12}$$

e, em seguida, a distribuição conjunta das unidades amostrais em R_m é dada por

$$f(\boldsymbol{Y}_m|\theta_m) = \left(2\pi\sigma^2\right)^{-n_m/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n_m}(y_{mj} - \mu_m)^2\right\},$$

em que $\theta_m = (\mu_m, \sigma^2)$. Observa-se que o modelo adotado em (3.12) considera que a variância é constante em cada região de $\boldsymbol{\mathcal{X}}$ e por essa característica é conhecido como modelo de regressão *mean-shift*. A função de verossimilhança para o modelo de regressão considerando o modelo de partição bayesiana é dada por

$$L(\boldsymbol{\mathcal{T}}, \boldsymbol{\theta}|\mathcal{D}) = \prod_{m=1}^{M} f(\boldsymbol{Y}_m|\theta_m), \tag{3.13}$$

em que $\boldsymbol{\theta} = (\mu_1, \ldots, \mu_m, \sigma^2)$. A fim de obter a verossimilhança marginal para a tesselação $\boldsymbol{\mathcal{T}}$ foram atribuídas distribuições *a priori* para os parâmetros locais. Nesse sentido, para a média da distribuição normal, em cada região assumiu-se uma distribuição normal, e para a variância, for considerada uma distribuição gama inversa

$$\begin{aligned}\mu_m &\sim \mathrm{N}(\mu_0, \sigma^2/v), \quad m = 1, \ldots, M\\ \sigma^2 &\sim \mathrm{IGa}(\sigma_0, \sigma_1),\end{aligned}$$

em que $\mu_0, v, \sigma_0, \sigma_1$ são hiperparâmetros especificados. Considerando-se as distribuições *a priori* definidas anteriormente para os parâmetros do modelo normal tem-se, em seguida, que a verossimilhança marginal para $\boldsymbol{\mathcal{T}}$ é dada por

$$L(\boldsymbol{\mathcal{T}}|\mathcal{D}) = \frac{cv^{M/2}}{\prod_{m=1}^{M}(n_m + v)^{1/2}}\left(0.5\left[\sum_{m=1}^{M}(n_m - 1)s_m^2 + \frac{n_m v}{n_m + v}(\bar{y}_m - \mu_0)^2\right] + \sigma_1\right)^{-(n/2+\sigma_0)}, \tag{3.14}$$

em que c é uma constante que não depende de $\boldsymbol{\mathcal{T}}$, $\bar{y}$ e s_m^2 representam a média amostral e a variância amostral na região R_m respectivamente.

A distribuição *a posteriori* para a tesselação $\boldsymbol{\mathcal{T}}$ é dada por

$$p(\boldsymbol{\mathcal{T}}|\mathcal{D}) \propto L(\boldsymbol{\mathcal{T}}|\mathcal{D})p(\boldsymbol{\mathcal{T}}),$$

em que a verossimilhança marginal $L(\boldsymbol{\mathcal{T}}|\mathcal{D})$ é dada em (3.14) e a distribuição *a priori* para $\boldsymbol{\mathcal{T}}$ foi definida em (3.5). Para simular amostras de $p(\boldsymbol{\mathcal{T}}|\mathcal{D})$ foi utilizada a estratégia computacional apresentada na Seção 3.1.3. As distribuições condicionais completas para μ_m e σ^2 são dados respectivamente, por

$$\mu_m|\boldsymbol{\mathcal{T}}, \sigma^2, \mathcal{D} \sim \mathrm{N}\left(\frac{n_m\bar{y}_m + v\mu_0}{n_m + v}, \frac{\sigma^2}{n_m + v}\right)$$

e

$$\sigma^2|\boldsymbol{\mathcal{T}}, \mathcal{D} \sim \mathrm{IGa}\left(\frac{n}{2} + \alpha, \frac{1}{2}\left[\sum_{m=1}^{M}(n_m - 1)s_m^2 + \frac{n_m v}{n_m + v}(\bar{y}_m - \mu_0)^2\right] + \sigma_1\right)$$

Exemplo 3.2. Em análise de sobrevivência, a distribuição exponencial é amplamente utilizada para modelar tempos de falha. Uma característica da distribuição exponencial esta baseada no fato que assume que os indivíduos têm um risco constante ao longo do tempo de estudo. Assim, supondo que uma variável aleatória não negativa T tem distribuição exponencial então a função densidade é dada por

$$f(t|\theta) = \theta\exp(-\theta t), \quad t > 0, \quad \theta > 0. \tag{3.15}$$

Denota-se por $T \sim \mathrm{Exp}(\theta)$ se T segue uma distribuição exponencial com parâmetro θ. A função de sobrevivência para T é dada por $S(t|\theta) = \exp(-\theta t)$.

Considerando-se uma extensão local para o modelo exponencial baseado no modelo MPB, supõe-se que a tesselação por hiperplanos divide em M regiões o espaço preditor, sendo que o parâmetro local em cada região é dada pelo parâmetro da distribuição exponencial.

Seja T_{mj} o tempo de falha para o j-ésimo indivíduo na região R_m e C_{mj} o tempo da censura. O tempo observado, é dado por $Y_{mj} = \min\{T_{mj}, C_{mj}\}$. A variável indicadora de censura δ_{mj} é definida sendo $\delta_{mj} = 1$ se $Y_{mj} = T_{mj}$, e $\delta_{mj} = 0$ caso contrário.

A função de verossimilhança para os dados considerando censura não informativa é dada por

$$L(\boldsymbol{\mathcal{T}}, \boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{\delta}) = \prod_{m=1}^{M}\prod_{j=1}^{n_m}\{f(y_{mj}|\theta_m)\}^{\delta_{mj}}\{S(y_{mj}|\theta_m)\}^{1-\delta_{mj}} = \prod_{m=1}^{M}\theta_m^{\nu_m}\exp(-\theta_m\sum_{j=1}^{n_m}y_{mj})$$

em que $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M)^\top$, $\nu_m = \sum_{j=1}^{n_m}\delta_{mj}$, $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ e $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^\top$.

A distribuição *a priori* conjugada para θ é uma distribuição gama, e assim, assume-se que essa seja a distribuição *a priori* para os parâmetros locais θ_m e tem-se

$$\theta_m \sim \text{Ga}(a_0, a_1) \quad m = 1, \ldots, M,$$

em que a_0, a_1 são parâmetros da distribuição gama. A distribuição *a posteriori* para $(\boldsymbol{\theta}, \boldsymbol{\mathcal{T}})$ é dada por

$$p(\boldsymbol{\mathcal{T}}, \boldsymbol{\theta}|\mathcal{D}) \propto L(\boldsymbol{\mathcal{T}}, \boldsymbol{\theta}|\mathcal{D}) p(\boldsymbol{\mathcal{T}}, \boldsymbol{\theta}).$$

Note-se que a distribuição *a posteriori* $p(\boldsymbol{\mathcal{T}}, \boldsymbol{\theta}|\mathcal{D})$ pode ser fatorizada na forma

$$p(\boldsymbol{\mathcal{T}}, \boldsymbol{\theta}|\mathcal{D}) = p(\boldsymbol{\theta}|\boldsymbol{\mathcal{T}}, \mathcal{D}) p(\boldsymbol{\mathcal{T}}|\mathcal{D}),$$

e tem-se que a distribuição condicional completa para θ_m é dada por

$$\theta_m|\boldsymbol{\mathcal{T}}, \mathcal{D} \sim \text{Ga}(\nu_m + a_0, \sum_{j=1}^{n_m} y_{mj} + a_1). \tag{3.16}$$

A distribuição *a posteriori* de $\boldsymbol{\mathcal{T}}$ é definida sendo $p(\boldsymbol{\mathcal{T}}|\mathcal{D}) \propto L(\boldsymbol{\mathcal{T}}|\mathcal{D}) p(\boldsymbol{\mathcal{T}})$ em que

$$L(\boldsymbol{\mathcal{T}}|\mathcal{D}) = \int L(\boldsymbol{\mathcal{T}}, \boldsymbol{\theta}|\mathcal{D}) p(\boldsymbol{\theta}|\boldsymbol{\mathcal{T}}) = \prod_{m=1}^{M} \frac{a_1^{a_0}}{\Gamma(a_0)} \frac{\Gamma(\nu_m + a_0)}{\left(\sum_{j=1}^{n_m} y_{mj} + a_1\right)^{(\nu_m + a_0)}}.$$

O tratamento analítico da distribuição *a posteriori* de $\boldsymbol{\mathcal{T}}$ é difícil, e por isso fez-se uso do amostrador MCMC proposto na Seção 3.1.3 para explorar $p(\boldsymbol{\mathcal{T}}|\mathcal{D})$.

3.3 Comentários finais

Neste capítulo, foi apresentado o modelo MPB considerando hiperplanos ortogonais. Estendeu-se a modelagem proposta por Holmes *et al.* (2005), por considerar no modelo de partição variáveis qualitativas (com mais de duas categorias) e, desta maneira, foi desenvolvida uma nova estrategia computacional para explorar a distribuição *a posteriori* da tesselação (veja a Seção 3.1.3).

Uma vantagem de considerar-se a tesselação por hiperplanos ortogonais para obter uma partição no espaço das covariáveis, $\boldsymbol{\mathcal{X}}$, é que os hiperplanos selecionam as covariáveis que têm influência no modelo considerando o criterio fator de Bayes. Nesse sentido, nota-se que se é proposto um ponto de corte em uma covariável obtem-se uma nova partição do espaço preditor em que essa nova partição é avaliada por meio do fator de Bayes (veja a equação (3.11)), assim caso a nova partição é aceita então significa que essa covariável é informativa no modelo e se for rejeitada significa que essa nova partição não é plausível e portanto essa variável não é influente no ajuste do modelo.

Capítulo 4

Modelagem local com partição bayesiana para o modelo de série de potências com fração de cura

Em geral, na presença de variáveis preditoras, os modelos de longa duração propostos na literatura fazem uso de uma função de ligação para relacionar as covariáveis com o parâmetro de fração de cura. Neste trabalho, foi usada uma estrutura local no espaço das covariáveis $\boldsymbol{\mathcal{X}}$ e, desta forma, os efeitos das covariáveis são capturados através de um modelo local. Para este fim foi utilizado o modelo de partição bayesiana proposto no Capítulo 3, porém aplicado para dados de sobrevivência com fração de cura.

Para a construção das regiões em $\boldsymbol{\mathcal{X}}$, foi adotada a tesselação por hiperplanos ortogonais como foi proposto na Seção 3.1. Uma vantagem de trabalhar com a tesselação por hiperplanos ortogonais é seleção das covariáveis que têm efeito na variável resposta. Neste trabalho, a estratégia computacional para trabalhar com hiperplanos ortogonais proposta por Hoggart & Griffin (2001) é modificada para considerar covariáveis qualitativas com mais de duas categorias.

Na Seção 4.1 é apresentado a extensão local do modelo de longa duração quando o número de causas latentes seguem uma distribuição de série de potências considerando o modelo de partição bayesiana. Na Seção 4.2 são apresentados alguns casos particulares da extensão local do modelo de longa duração de série de potências; na Seção 4.3, será descrito o critério de seleção de modelo para os modelos de longa duração com partição e na Seção 4.4, serão mostradas 2 aplicações a dados reais.

4.1 Modelagem local por hiperplanos ortogonais

A tesselação por hiperplanos ortogonais $\boldsymbol{\mathcal{T}}$ define M regiões $R_1, \ldots, R_M$ no espaço preditor $\boldsymbol{\mathcal{X}}$ e seja n_m o número de observações na região R_m.

Denote-se por N_{mj} (não observável) o número de causas do evento de interesse da j-ésima observação na m-ésima região, R_m, com distribuição de probabilidade $p(N_{mj}|\theta_m)$, $j = 1, \ldots, n_m$.

Dado N_{mj}, sejam $Z_{mj}^1, \ldots, Z_{mj}^{N_{mj}}$ os tempos de ocorrência do evento de interesse para N_{mj}, com função de distribuição acumulada $F(\cdot) = 1 - S(\cdot)$, em que $S(\cdot)$ é função de sobrevivência. Neste trabalho, considera-se uma forma paramétrica para a $F(\cdot)$, como por exemplo, a distribuição Weibull, gama generalizada, Gompertz entre outras. A distribuição acumulada foi indexada pelo vetor de parâmetros $\boldsymbol{\gamma}$, $F(\cdot) = F(\cdot|\boldsymbol{\gamma})$.

Seja T_{mj} como em (2.1) e C_{mj} o tempo da censura. O tempo observado, é dado por $Y_{mj} = \min\{T_{mj}, C_{mj}\}$. Seja δ_{mj} a variável indicadora de censura, com $\delta_{mj} = 1$ se $Y_{mj} = T_{mj}$, e $\delta_{mj} = 0$ caso contrário.

A função de verossimilhança para os dados completos e considerando censura não informativa é dada por

$$L(\boldsymbol{\mathcal{T}}, \boldsymbol{\theta}, \boldsymbol{\gamma}|\boldsymbol{N}, \boldsymbol{y}, \boldsymbol{\delta}) = \prod_{m=1}^{M} \prod_{j=1}^{n_m} \{S(y_{mj}|\boldsymbol{\gamma})\}^{N_{mj}-\delta_{mj}} \{N_{mj} f(y_{mj}|\boldsymbol{\gamma})\}^{\delta_{mj}} p(N_{mj}|\theta_m), \tag{4.1}$$

em que $\boldsymbol{N} = (N_1, \ldots, N_n)^\top$ é o vetor de variáveis latentes, $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M)^\top$ e $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^\top$.

Observe-se que, em cada região R_m, o número de causas para o evento de interesse N_{mj} tem a mesma distribuição de probabilidade com parâmetro local θ_m, $p(N_{mj}|\theta_m)$, portanto a distribuição para o número de causas ou riscos sob a tesselação $\boldsymbol{\mathcal{T}}$ e parâmetros locais $\boldsymbol{\theta}$ é dada por

$$p(\boldsymbol{N}|\boldsymbol{\theta}, \boldsymbol{\mathcal{T}}) = \prod_{m=1}^{M} p(\boldsymbol{N}_m|\theta_m) = \prod_{m=1}^{M} \prod_{j=1}^{n_m} p(N_{mj}|\theta_m), \tag{4.2}$$

em que $\boldsymbol{N}_m = (N_{m1}, \ldots, N_{mn_m})$, $m = 1, \ldots, M$.

O número de causas latentes N segue uma distribuição de série de potências com distribuição de probabilidade definida em (2.6), e os casos particulares estão apresentados na Tabela 2.2. Também, assume-se a distribuição Weibull para o tempo de ocorrência Z_{mj} considerando a parametrização como em Ibrahim *et al.* (2001b) dada em (2.32). A parametrização dada em (2.32) permite uso do algoritmo rejeição adaptativa (Gilks & Wild, 1992).

4.1.1 Análise bayesiana

Considerando-se a metodologia do modelo de partição bayesiana, a distribuição *a priori* conjunta para $(\gamma, \theta, \boldsymbol{\mathcal{T}})$ é dada por

$$p(\boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\mathcal{T}}) = p(\boldsymbol{\gamma})p(\boldsymbol{\theta}, \boldsymbol{\mathcal{T}}) = p(\boldsymbol{\gamma})p(\boldsymbol{\theta}|\boldsymbol{\mathcal{T}})p(\boldsymbol{\mathcal{T}}).$$

Tendo-se assumido que os parâmetros da distribuição Weibull são independentes, tem-se que $p(\boldsymbol{\gamma}) = p(\alpha)p(\lambda)$ em que $\alpha \sim \text{Ga}(\mu_\alpha, \sigma_\alpha)$ e $\lambda \sim \text{N}(\mu_\lambda, \sigma_\lambda)$, sendo que μ_α, σ_α, μ_λ e σ_λ são hiperparâmetros.

O modelo de partição bayesiana considera que os parâmetros locais entre as regiões R_m são independentes, assim a distribuição *a priori* para $\boldsymbol{\theta}$ é dada por

$$p(\boldsymbol{\theta}|\boldsymbol{\mathcal{T}}) = \prod_{m=1}^{M} p(\theta_m|\boldsymbol{\mathcal{T}}),$$

em que $p(\theta_m|\boldsymbol{\mathcal{T}})$ é a distribuição *a priori* para θ_m.

Foi introduzido o vetor $\boldsymbol{N}$ de variáveis latentes para obter as amostras da distribuição *a posteriori* $p(\boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\mathcal{T}}|\boldsymbol{y}, \boldsymbol{\delta})$. Desse modo, a distribuição *a posteriori* conjunta $p(\boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\mathcal{T}}, \boldsymbol{N}|\boldsymbol{y}, \boldsymbol{\delta})$ é dada por

$$\begin{aligned} p(\boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\mathcal{T}}, \boldsymbol{N}|\boldsymbol{y}, \boldsymbol{\delta}) \propto \prod_{m=1}^{M} \exp\left\{-e^\lambda \sum_{j=1}^{n_m} y_{mj}^\alpha N_{mj}\right\} \prod_{j=1}^{n_m} \left(N_{mj}\alpha e^\lambda y_{mj}^{\alpha-1}\right)^{\delta_{mj}} p(N_{mj}|\theta_m) \\ \times p(\boldsymbol{\gamma})p(\boldsymbol{\theta}, \boldsymbol{\mathcal{T}}). \end{aligned} \tag{4.3}$$

A distribuição *a posteriori* dos parâmetros do modelo não têm uma forma analítica, portanto foram usados métodos computacionais MCMC (Brooks *et al.*, 2011) para simular amostras da distribuição *a posteriori*. Em seguida, obtem-se amostras das condicionais completas $(\boldsymbol{\theta}, \boldsymbol{\mathcal{T}}|\boldsymbol{N}, \boldsymbol{\gamma}, \boldsymbol{y}, \boldsymbol{\delta})$, $(\boldsymbol{N}|\boldsymbol{\theta}, \boldsymbol{\mathcal{T}}, \boldsymbol{\gamma}, \boldsymbol{y}, \boldsymbol{\delta})$ e $(\boldsymbol{\gamma}|\boldsymbol{\theta}, \boldsymbol{\mathcal{T}}, \boldsymbol{N}, \boldsymbol{y}, \boldsymbol{\delta})$.

Supondo-se que os parâmetros da distribuição Weibull são independentes, tem-se que as distribuições condicionais completas são expressas por

$$\begin{aligned} p(\lambda|\alpha, \boldsymbol{N}, \boldsymbol{\mathcal{T}}, \boldsymbol{y}, \boldsymbol{\delta}) &\propto e^{d\lambda} \exp\left(-e^\lambda \sum_{i=1}^{n} N_i y_i^\alpha\right) \exp\left(-\frac{(\lambda - \mu_\lambda)^2}{2\sigma_\lambda^2}\right) \\ p(\alpha|\lambda, \boldsymbol{N}, \boldsymbol{\mathcal{T}}, \boldsymbol{y}, \boldsymbol{\delta}) &\propto \alpha^d \left(\prod_{i=1}^{n} y_i^{\delta_i}\right)^\alpha \exp\left(-e^\lambda \sum_{i=1}^{n} N_i y_i^\alpha\right) \alpha^{\mu_\alpha - 1} e^{-\sigma_\alpha \alpha} \end{aligned}$$

em que $d = \sum_{i=1}^{n} \delta_i$.

Por outro lado, observe-se que para simular amostras de $(\boldsymbol{\theta}, \boldsymbol{\mathcal{T}}|\boldsymbol{\gamma}, \boldsymbol{y}, \boldsymbol{\delta})$, foi considerada a condicional completa dada por

$$p(\boldsymbol{\theta}, \boldsymbol{\mathcal{T}}|\boldsymbol{N}, \boldsymbol{\gamma}, \boldsymbol{y}, \boldsymbol{\delta}) = p(\boldsymbol{\mathcal{T}}|\boldsymbol{N}, \boldsymbol{\gamma}, \boldsymbol{y}, \boldsymbol{\delta})p(\boldsymbol{\theta}|\boldsymbol{\mathcal{T}}, \boldsymbol{N}, \boldsymbol{\gamma}, \boldsymbol{y}, \boldsymbol{\delta}),$$

e, dessa maneira, a distribuição condicional completa para $(\boldsymbol{\mathcal{T}}|\boldsymbol{N}, \boldsymbol{\gamma}, \boldsymbol{y}, \boldsymbol{\delta})$ é dada por

$$p(\boldsymbol{\mathcal{T}}|\boldsymbol{N}, \boldsymbol{\gamma}, \boldsymbol{y}, \boldsymbol{\delta}) \propto p(\boldsymbol{N}|\boldsymbol{\mathcal{T}})p(\boldsymbol{\mathcal{T}}),$$

em que

$$p(\boldsymbol{N}|\boldsymbol{\mathcal{T}}) = \int p(\boldsymbol{N}|\boldsymbol{\theta}, \boldsymbol{\mathcal{T}})p(\boldsymbol{\theta}|\boldsymbol{\mathcal{T}})\mathrm{d}\boldsymbol{\theta}. \tag{4.4}$$

Para melhorar a convergência e o *mixing* do amostrador de Gibbs (Chen *et al.*, 2000) foi integrado o vetor de parâmetros locais $\boldsymbol{\theta}$ em (4.4). A técnica anterior é conhecida como amostrador de Gibbs por colapso (do inglês, Collapsed Gibbs Sampler) (Liu, 1994).

Neste sentido, a integral dada em (4.4) pode ter uma uma forma fechada e, para tal fim, foi necessário atribuir distribuições *a priori* para os parâmetros locais θ_m, de forma que seja possível o tratamento analítico de (4.4). Se o número N de riscos latentes segue a distribuição de série de potências, tem-se que para diferentes funções de série, obtém-se diferentes distribuições de probabilidade e, dependendo dessa distribuição, escolhe-se uma distribuição *a priori* para θ_m de maneira que a integração em (4.4) seja feita analiticamente. Na seguinte seção, serão apresentados os casos particulares que serão desenvolvidos ao longo deste capítulo.

4.2 Casos Particulares

4.2.1 Modelo de fração de cura binomial com partição bayesiana (MPBBi)

Considerando, em cada região R_m, que o número de causas latentes para o evento de interesse N_{mj} segue a distribuição binomial com parâmetros K e θ_m, $N_{mj} \sim \mathrm{Bi}(K, \theta_m)$, tem-se que a distribuição de probabilidade é dada por

$$p(\boldsymbol{N}_{mj}|\theta_m, \boldsymbol{\mathcal{T}}) = \binom{K}{N_{mj}} \theta_m^{N_{mj}} (1-\theta_m)^{K-N_{mj}}, \quad N_{mj} = 1, \ldots, K,$$

a distribuição para o número de causas ou riscos sob a tesselação $\boldsymbol{\mathcal{T}}$ e os parâmetros locais $\boldsymbol{\theta}$ é dada por

$$p(\boldsymbol{N}|\boldsymbol{\theta}, \boldsymbol{\mathcal{T}}) = \prod_{m=1}^{M} \prod_{j=1}^{n_m} \binom{K}{N_{mj}} \theta_m^{N_{mj}} (1-\theta_m)^{K-N_{mj}}.$$

A fim de obter uma forma explícita para $p(\boldsymbol{N}|\boldsymbol{\mathcal{T}})$, atribui-se uma distribuição *a priori* para os parâmetros locais θ_m, de forma que a integral em (4.4) seja analiticamente tratável.

Por isso, adota-se uma distribuição beta como distribuição *a priori* em cada região, pelo fato dessa distribuição ser uma distribuição *a priori* conjugada para o parâmetro θ_m

$$\theta_m|\boldsymbol{\mathcal{T}} \sim \text{Be}(a_0, a_1), \quad m = 1, \ldots, M,$$

em que a_0 e a_1 são os hiperparâmetros especificados. Em seguida, tem-se uma forma explícita para a expressão (4.4) dada por

$$p(\boldsymbol{N}|\boldsymbol{\mathcal{T}}) = \prod_{i=1}^{n} \binom{K}{N_i} \prod_{m=1}^{M} \frac{\mathcal{B}(\sum_{j=1}^{n_m} N_{mj} + a_0, Kn_m - \sum_{j=1}^{n_m} N_{mj} + a_1)}{\mathcal{B}(a_0, a_1)},$$

em que $\mathcal{B}(\cdot, \cdot)$ é a função beta.

A distribuição condicional completa completa para θ_m é dada por

$$\theta_m|\boldsymbol{N}, \boldsymbol{\mathcal{T}} \sim \text{Be}\left(\sum_{j=1}^{n_m} N_{mj} + a_0, Kn_m - \sum_{j=1}^{n_m} N_{mj} + a_1\right), \quad m = 1, \ldots, M.$$

A distribuição condicional completa para as causas latentes N_{mj}'s é dada por

$$N_{mj}|\boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\mathcal{T}}, \boldsymbol{y}, \boldsymbol{\delta} \sim \text{Bi}\left(K - \delta_{mj}, \frac{S(y_{mj}, \gamma)\theta_m}{1 - \theta_m}\right) + \delta_{mj}.$$

Neste trabalho, assumiu-se que o parâmetro K da distribuição binomial é fixado.

4.2.2 Modelo de fração de cura Poisson com partição bayesiana (MPBPoi)

Considerando que em cada região R_m do espaço preditor o número de causas para o evento de interesse N_{mj} segue uma distribuição Poisson com distribuição de probabilidade dada por

$$p(\boldsymbol{N}_{mj}|\theta_m, \boldsymbol{\mathcal{T}}) = \frac{e^{-\theta_m}\theta_m^{N_{mj}}}{N_{mj}!},$$

tem-se que modelo para o número de causas ou riscos sob a tesselação $\boldsymbol{\mathcal{T}}$ e $\boldsymbol{\theta}$ é dada por

$$p(\boldsymbol{N}|\boldsymbol{\theta}, \boldsymbol{\mathcal{T}}) = \prod_{m=1}^{M} \prod_{j=1}^{n_m} \frac{e^{-\theta_m}\theta_m^{N_{mj}}}{N_{mj}!}.$$

Considera-se uma distribuição gama como a distribuição *a priori* para o parâmetro local θ_m,

$$\theta_m|\boldsymbol{\mathcal{T}} \sim \text{Ga}(b_0, b_1), \quad m = 1, \ldots, M,$$

em que b_0 e b_1 são hiperpârametros especificados. Observe-se que a distribuição gama é uma distribuição *a priori* conjugada para θ_n e, em seguida, obtém-se uma expressão fechada para (4.4) dada por

$$p(\boldsymbol{N}|\boldsymbol{\mathcal{T}}) = \prod_{m=1}^{M} \frac{1}{\prod_{j=1}^{n_m} N_{mj}!} \frac{b_1^{b_0}}{\Gamma(b_1)} \frac{\Gamma(\sum_{j=1}^{n_m} N_{mj} + b_0)}{(n_m + b_1)^{\sum_{j=1}^{n_m} N_{mj} + b_0}} \tag{4.5}$$

A distribuição condicional completa para o parâmetro θ_m é dada por

$$\theta_m|\boldsymbol{N}, \boldsymbol{\mathcal{T}} \sim \text{Ga}\left(\sum_{j=1}^{n_m} N_{mj} + b_0, n_m + b_1\right),$$

e a distribuição condicional para o número de causas latentes é dada por

$$N_{mj}|\boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\mathcal{T}}, \boldsymbol{y}, \boldsymbol{\delta} \sim \text{Poi}\left(\theta_m S(y_{mj}|\gamma)\right) + \delta_{mj}.$$

4.2.3 Modelo de fração de cura binomial negativa com partição bayesiana (MPBBn)

Assumindo-se que o número de causas latentes N_{mj} em cada região segue uma distribuição binomial negativa, tem-se que a função de probabilidade é dada por

$$p(N_{mj}|\theta_m) = \binom{\tau + N_{mj} - 1}{\tau - 1} \theta_m^{N_{mj}} (1 - \theta_m)^{\tau}, \quad N_{mj} = 0, 1, 2, \ldots, \quad 0 < \theta < 1, \tag{4.6}$$

em que τ é um inteiro positivo. A média e variância são, respectivamente, $\text{E}[N_{mj}] = \tau\theta_m/(1-\theta_m)$ e $\text{Var}[N_{mj}] = \tau\theta_m/(1-\theta_m)^2$. Considerando a distribuição binomial negativa dada em (4.6) tem-se que a distribuição conjunta para o vetor de riscos latentes levando em conta a tesselação $\boldsymbol{\mathcal{T}}$ é dada por,

$$p(\boldsymbol{N}|\boldsymbol{\theta}, \boldsymbol{\mathcal{T}}) = \prod_{m=1}^{M} \prod_{j=1}^{n_m} \binom{\tau + N_{mj} - 1}{\tau - 1} \theta_m^{N_{mj}} (1 - \theta_m)^{\tau},$$

sendo que neste trabalho foi fixado o parâmetro τ para diferentes valores.

Adota-se a distribuição beta como distribuição *a priori* para θ_m

$$\theta_m|\boldsymbol{\mathcal{T}} \sim \text{Be}(c_0, c_1), \quad m = 1, \ldots, M,$$

em que c_0 e c_1 são hiperparâmetros especificados. Considerando que a distribuição beta é uma distribuição conjugada para θ_m, a integral dada em (4.4) é dada por

$$p(\boldsymbol{N}|\boldsymbol{\mathcal{T}}) = \prod_{i=1}^{n} \binom{\tau + N_i - 1}{\tau - 1} \prod_{m=1}^{M} \frac{\mathcal{B}(\tau n_m + c_0, \sum_{j=1}^{n_m} N_{mj} + c_1)}{\mathcal{B}(c_0, c_1)},$$

em que $\mathcal{B}(\cdot,\cdot)$ é a função beta.

A distribuição condicional completa para o parâmetro θ_m é dada por

$$\theta_m|\boldsymbol{\mathcal{T}},\boldsymbol{N} \sim \text{Be}\left(\sum_{j=1}^{n_m} N_{mj} + c_0, \tau n_m + c_1\right).$$

Por outro lado, sabe-se que N_{mj}'s são variáveis independentes, então a distribuição condicional completa para N_{mj} em cada região R_m é dada por

$$N_{mj}|\boldsymbol{\gamma},\boldsymbol{\theta},\boldsymbol{\mathcal{T}},\boldsymbol{y},\boldsymbol{\delta} \sim \text{Bn}\left(\tau+\delta_{mj}, \theta_m \exp\left(-e^{\lambda} y_{mj}^{\alpha}\right)\right) + \delta_{mj}.$$

Observe-se que se $\tau = 1$, obtém-se o modelo de fração de cura geométrica com partição bayesiana (MPBGeo).

4.2.4 Modelo de fração de cura logarítmica com partição bayesiana (MPBLg)

Assumiu-se que o número de riscos latentes N_{mj} segue uma distribuição logarítmica com distribuição de probabilidade dada por

$$p(N_{mj}|\theta_m) = \frac{\theta_m^{N_{mj}+1}}{-(N_{mj}+1)\log(1-\theta_m)} \quad N_{mj} = 0,1,\ldots, \quad 0<\theta_m<1. \tag{4.7}$$

A distribuição conjunta para $\boldsymbol{N}$ sob a tesselação $\boldsymbol{\mathcal{T}}$ e parâmetros locais $\boldsymbol{\theta}$ é dada por

$$p(\boldsymbol{N}|\boldsymbol{\theta},\boldsymbol{\mathcal{T}}) = \prod_{m=1}^{M}\prod_{j=1}^{n_m} \frac{\theta_m^{N_{mj}+1}}{-(N_{mj}+1)\log(1-\theta_m)}.$$

Diferente das outras distribuições desenvolvidas anteriormente, a distribuição logarítmica não tem uma distribuição *a priori* conjugada. Porém, foi considerada como distribuição *a priori* para θ_m uma distribuição beta

$$\theta_m|\boldsymbol{\mathcal{T}} \sim \text{Be}(d_0,d_1), \quad m=1,\ldots,M, \tag{4.8}$$

em que d_0 e d_1 são hiperparâmetros especificados.

Pelo fato da distribuição logarítmica não ter uma distribuição *a priori* conjugada, a integral dada em (4.4) não pode ser encontrada analiticamente, portanto foi usada integração numérica, que será aplicada em cada região R_m

$$p(\boldsymbol{N}_m|\boldsymbol{\mathcal{T}}) = \frac{1}{\mathcal{B}(d_0,d_1)}\frac{1}{\prod_{j=1}^{n_m}(N_{mj}+1)}\int_0^1 \frac{\theta_m^{n_m+\sum N_{mj}+d_0-1}(1-\theta_m)^{d_1-1}}{\{-\log(1-\theta_m)\}^{n_m}}\ \text{d}\theta_m \quad m=1,\ldots,M, \tag{4.9}$$

em que $\mathcal{B}(\cdot,\cdot)$ é a função beta. Assim, a distribuição condicional completa de θ_m é dada por

$$p(\theta_m|\boldsymbol{N},\boldsymbol{\mathcal{T}}) \propto \frac{\theta_m^{n_m+N_{mj}+d_0-1}(1-\theta_m)^{d_1-1}}{\{-\log(1-\theta_m)\}^{n_m}},$$

em que, para obter amostras desta distribuição condicional, será usado o método de rejeição adapativo (Gilks & Wild, 1992).

Considerando a suposição que os N_{mj}'s são variáveis independentes, a distribuição condicional completa de N_{mj} em cada região R_m é dada por

$$p(N_{mj}|\boldsymbol{\gamma},\boldsymbol{\theta},\boldsymbol{\mathcal{T}},\boldsymbol{y},\boldsymbol{\delta}) \propto \exp\left(-e^{\lambda}y_{mj}^{\alpha}N_{mj}\right) N_{mj}^{\delta_{mj}} \frac{\theta_m^{N_{mj}}}{N_{mj}+1}. \tag{4.10}$$

No caso em que $\delta_{mj}=0$, a distribuição condicional *a posteriori* para o número de causas do evento de interesse N_{mj} é dada por

$$N_{mj}|\boldsymbol{\gamma},\boldsymbol{\theta},\boldsymbol{\mathcal{T}},\boldsymbol{y},\boldsymbol{\delta} \sim \mathrm{Lg}(\theta_m S(y_{mj}|\boldsymbol{\gamma})),$$

não obstante se $\delta_{mj}=1$ tem-se que

$$p(N_{mj}|\boldsymbol{\gamma},\boldsymbol{\theta},\boldsymbol{\mathcal{T}},\boldsymbol{y},\boldsymbol{\delta}) \propto \frac{N}{N+1}\left\{\theta_m S(y_{mj}|\boldsymbol{\gamma})\right\}^N.$$

Para a geração de N_{mj}, considera-se o algoritmo proposto em Kemp (1981), adaptado ao caso em que a distribuição logarítmica é deslocada no zero.

Nos modelos de longa duração com partição bayesiana propostos anteriormente foi considerado distribuições *a priori* conjugadas para o parâmetros locais. Porém no caso da distribuição logarítmica não existe uma distribuição conjugada e desta forma foi utilizado integração numérica para calcular a integral dada em (4.4). Nesse sentido, pode-se considerar outras distribuições *a priori* para as outras distribuições, por exemplo, pode-se atribuir uma distribuição log-normal como sendo uma distribuição *a priori* para o parâmetro da distribuição Poisson. Não obstante, o custo computacional para calcular a integral em (4.4) aumenta.

4.3 Comparação de modelos

Para avaliar a qualidade do ajuste do modelo aos dados, foi considerada a densidade preditiva condicional ordinária (CPO) (Ibrahim *et al.*, 2001b).

Seja $\mathcal{D}^{(-i)}$ que denota os dados com a i-ésima observação excluída. Para cada modelo proposto, ficou definida $g(y_i|\boldsymbol{\vartheta}) = S_{pop}(y_i|\boldsymbol{\vartheta})$ para os tempos observados ($\delta_i = 1$) e

$g(y_i|\boldsymbol{\vartheta}) = f_{pop}(y_i|\boldsymbol{\vartheta})$ para os tempos censurados ($\delta_i = 0$) em que $\boldsymbol{\vartheta} = (\boldsymbol{\theta}, \boldsymbol{\gamma})^\top$. Foi denotada a densidade *a posteriori* de $\boldsymbol{\vartheta}$ dado $\mathcal{D}^{-i} = (y_i, \delta_i)$ por $p(\boldsymbol{\vartheta}|\mathcal{D}^{-i})$, $i = 1, \ldots, n$, logo CPO_i para a i-ésima observação é dada por

$$\text{CPO}_i = \int g(y_i|\boldsymbol{\vartheta}) p(\boldsymbol{\vartheta}|\mathcal{D}^{-i}) \text{d}\boldsymbol{\vartheta} = \left\{ \int \frac{p(\boldsymbol{\vartheta}|\mathcal{D})}{g(y_i|\boldsymbol{\vartheta})} \text{d}\boldsymbol{\vartheta} \right\}^{-1}. \tag{4.11}$$

Valores altos de CPO_i implicam um bom ajuste do modelo. Porém não ficou estabelecida uma forma fechada para CPO_i. Para estimar (4.11) utilizam-se as amostras MCMC da distribuição *a posteriori* $p(\boldsymbol{\vartheta}|\mathcal{D})$, portanto uma estimativa Monte Carlo para CPO_i (Chen *et al.*, 2000) é dada por

$$\widehat{\text{CPO}_i} = \left\{ \frac{1}{B} \sum_{b=1}^{B} \frac{1}{g(y_i|\boldsymbol{\vartheta}_i)} \right\}^{-1},$$

em que B denota o tamanho da amostra MCMC. Baseada nos CPO_i's outra medida para comparação é a estatística definida por $\text{LPML} = \sum_{b=1}^{B} \log(\text{CPO}_i)$, em que um valor alto de LPML indica um melhor ajuste do modelo considerado.

4.4 Aplicação

4.4.1 Dados de melanoma

Considerando os dados do Exemplo 2.2, apresentado no Capítulo 2, foi aplicado a metodologia proposta nesta seção. Para as estimativas bayesianas dos parâmetros da distribuição Weibull adotaram-se as distribuições *a priori* como foi visto na Seção 4.1.1, em que $\alpha \sim \text{Ga}(0,1; 0,1)$ e $\lambda \sim \text{N}(0, 100)$. Para o número de regiões M na tesselação assumimos uma distribuição geométrica com média 10, $M \sim \text{Geo}(0.1)$.

Na simulação MCMC, foram geradas duas cadeias independentes com 700000 iterações para os modelos de longa duração com partição bayesiana propostos na Seção 4.2. As primeiras 300000 foram descartadas como iterações *burn-in*, e foi adotado um salto de tamanho 100, conduzindo a uma amostra final de tamanho 4000 para cada cadeia de cada caso. No começo do algoritmo consideramos, $\boldsymbol{N} = (1, \ldots, 1)$ e $M = 1$.

Para monitorar a convergência dos modelos ajustados nesta seção, foram consideradas as duas cadeias geradas pelo amostrador MCMC, por ter-se verificado a probabilidade de corte das variáveis e também a probabilidade *a posteriori* das partições de x_3. Finalmente,

a convergência dos parâmetros da distribuição Weibull foi monitorada com o auxilio do fator de redução de escala ($\widehat{R}$) proposto por Gelman & Rubin (1992).

Na tesselação por hiperplanos ortogonais, as variáveis x_1, x_4, x_5 são divididas no máximo, em 2 grupos pelo fato de que essas variáveis são binarias. Porém as covariáveis x_2 e x_6 são variáveis contínuas, por isso os hiperplanos dividem essas variáveis de acordo com os pontos de corte relacionados à distribuição marginal para cada variável. No caso em que essas variáveis são informativas para o modelo significa que elas são divididas pelo menos por um hiperplano.

Resultados para o modelo MPBBi

Caso o número de causas latentes N siga a distribuição binomial tem-se que a distribuição *a priori* para θ_m em cada região R_m é uma distribuição beta com parâmetros a_0 e a_1 e para essa aplicação assumidos iguais a 1. Além disso, o parâmetro K da distribuição binomial é fixo e, por isso, considera-se que o conjunto de valores para K é dado por $\{1, 2, 7, 10\}$.

Na Tabela 4.1 é apresentada a probabilidade de corte de cada uma das covariáveis para diferentes valores de K. Nota-se que a covariável x_3 (categoria do nódulo) tem uma alta probabilidade *a posteriori* (próximo de 1) de ser dividida e desta forma x_3 sempre é dividida, pelo menos, por um hiperplano. Além disso, ressalta-se o fato que a probabilidade de corte de x_3 não muda, se forem assumidos diferentes valores para K. Em seguida, pode-se afirmar que a covariável x_3 tem um efeito significativo no modelo e, portanto, na fração de cura.

Note-se que também a probabilidade de corte da covariável x_2 (idade) muda de acordo com os diferentes valores de K, em seguida , independente do valor assumido para K essa variável tem pouca influência no modelo devido ao fato de sua probabilidade ser baixa. As probabilidade de corte, para as outras covariáveis são próximas de zero, por isso elas não são informativas no modelo.

A variável categoria do nódulo (x_3) é uma variável qualitativa ordinal, com mais de duas categorias, e divisão (partição) desta covariável é feita considerando-se a ordem das categorias. As diferentes partições para x_3 foram apresentadas na Tabela 3.1, em seguida, considerando-se o amostrador MCMC proposto na Seção 3.1.3 são apresentadas as probabilidades *a posteriori* para cada partição de x_3 na Tabela 4.2.

Nota-se que as partições $\{1, 2, 3\}, \{4\}$ e $\{1, 2\}, \{3, 4\}$ na Tabela 4.2 se destacam devido ao fato de suas probabilidades *a posteriori* serem maiores em relação às demais. De acordo com os resultados obtidos, existe uma conexão entre o parâmetro K e a partição

Tabela 4.1: Probabilidade de corte para as covariáveis do conjunto de dados de melanoma considerando o modelo MPBBi.

	Variáveis					
K	x_1	x_2	x_3	x_4	x_5	x_6
1	0,016	0,258	0, 997	0,024	0,027	0,090
2	0,013	0,326	0,999	0,014	0,019	0,054
7	0,002	0,149	0, 999	0,003	0,003	0,032
10	0,001	0,123	0, 999	0,001	0,002	0,020

$\{1, 2, 3\}, \{4\}$, isto é, à medida que K cresce a probabilidade *a posteriori* dessa partição também cresce, o que não acontece com outros grupos.

Tabela 4.2: Probabilidade *a posteriori* para as partições da covariável x_3 considerando os dados de melanoma para o modelo MPBBi.

Partições	Probabilidade *a posteriori*			
	$K = 1$	$K = 2$	$K = 7$	$K = 10$
{1,2,3,4}	0.000	0.000	0.000	0.000
{1},{2,3,4}	0.005	0.001	0.001	0.001
$\{1, 2\}, \{3, 4\}$	0.148	0.132	0.127	0.123
$\{1, 2, 3\}, \{4\}$	0.639	0.692	0.807	0.830
{1},{2},{3,4}	0.011	0.006	0.001	0.001
{1},{2,3},{4}	0.085	0.074	0.028	0.020
{1,2},{3},{4}	0.091	0.084	0.035	0.025
{1},{2},{3},{4}	0.021	0.010	0.001	0.001

Na Tabela 4.3 é apresentado o critério LPML para os valores assumidos do parâmetro K do modelo MPBBi. O critério LPML indica que o modelo MPBBi com $K = 10$ tem um melhor ajuste em relação ao modelo MPBBi com os outros valores assumidos para o parâmetro K.

Tabela 4.3: Critério LPML para os modelos MPBBi.

	$K = 1$	$K = 2$	$K = 7$	$K = 10$
LPML	-525,154	-523,393	-522,036	-521,775

Considerando o modelo MPBBi com $K = 10$, na Figura 4.1(a) e 4.1(b) é apresentada a

evolução da probabilidade de corte das covariáveis x_2, x_3 e x_6 ao longo das iterações para cadeia 1 e cadeia 2, respectivamente. Também a Figura 4.1(c) mostra a probabilidade *a posteriori* do número de regiões da tesselação. A maior probabilidade *a posteriori* do número de regiões é quando $M = 2$. Intuitivamente, o fato anterior pode ser analisado de acordo com a Tabela 4.2, onde os agrupamentos que têm maior probabilidade *a posteriori* para a covariável x_3 são aqueles que tem 2 grupos. Sendo que os hiperplanos selecionam x_3 por ter uma influência significativa no modelo, é razoável pensar que, na maioria das vezes, o espaço preditor seja dividido em duas regiões.

Resultados para o modelo MPBPoi

No caso que N segue a distribuição Poisson, foi adotada uma distribuição gama para os parâmetros locais, em que os hiperparâmetros são dadas por $b_0 = b_1 = 0.1$

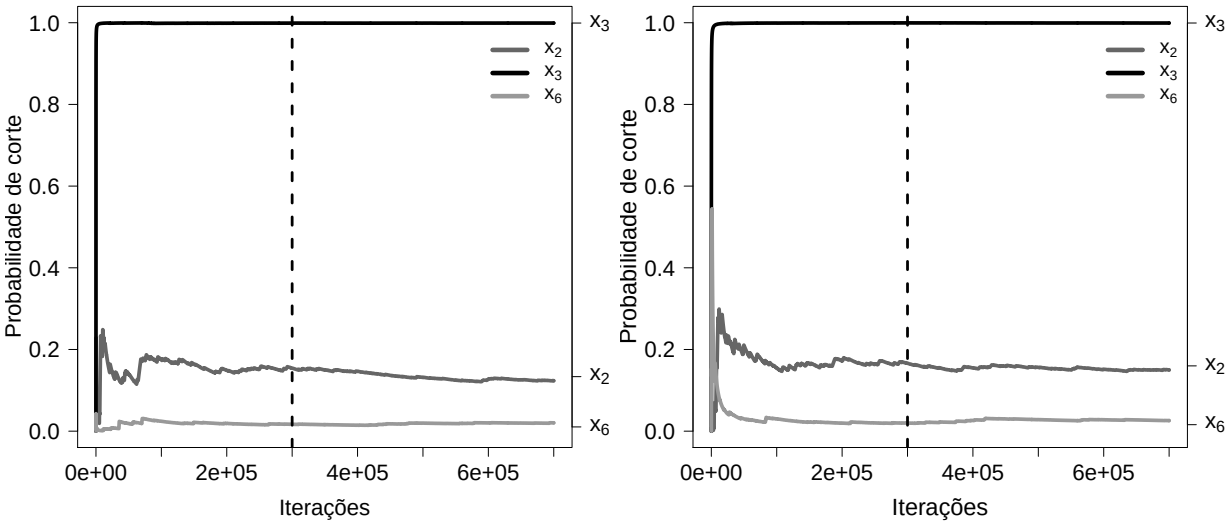
A Tabela 4.4 mostra as probabilidades de corte das covariáveis. As probabilidades de corte da covariável x_3 são próximas de 1. Isto significa que essa covariável faz parte da tesselação em grande parte das simulações MCMC e, portanto, a tesselação por hiperplanos ortogonais considera que x_3 tem um efeito significativo no modelo. Além disso, a probabilidade de corte da variável idade (x_2) é relativamente baixa e, em seguida, essa variável tem um efeito menor no modelo.

Tabela 4.4: Probabilidade de corte para as covariáveis do conjunto de dados de melanoma considerando o modelo MPBPoi.

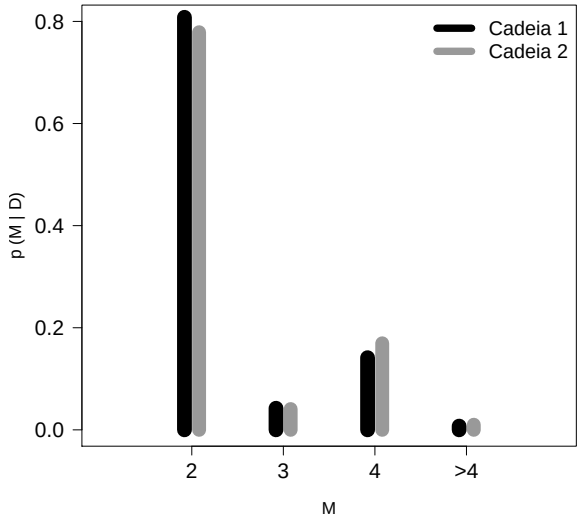
	Variáveis					
	x_1	x_2	x_3	x_4	x_5	x_6
Cadeia 1	0,018	0,307	1,000	0,025	0,024	0,102
Cadeia 2	0,019	0,299	0,998	0,027	0,023	0,096
Média	0,018	0,303	0,999	0,026	0,024	0,099

A tesselação por hiperplanos ortogonais mostra que x_3 é uma variável que tem efeito na fração de cura. A Tabela 4.5 apresenta as probabilidades *a posteriori* dos diferentes agrupamentos de x_3. Em seguida, similar ao modelo MPBBi, o agrupamento $\{1, 2, 3\}, \{4\}$ tem a maior probabilidade em relação aos outros agrupamentos de x_3

A Figura 4.2(a) e 4.2(b) apresenta a evolução da probabilidade de corte das covariáveis x_2, x_3 e x_6 ao longo da simulação MCMC para cadeia 1 e cadeia 2 respectivamente. A Figura 4.2(c) mostra a probabilidade *a posteriori* do número de regiões na tesselação. No modelo MPBPoi o número de regiões M com maior probabilidade é quando $M = 2$.



(a) (b)



(c)

Figura 4.1: Evolução da probabilidade corte na (a) cadeia 1 e (b) cadeia 2. (c) Probabilidade *a posteriori* do número de regiões, para os dados de melanoma seguindo o modelo MPBBi com $K = 10$.

Resultados para o modelo MPBBn

No caso em que o número de riscos latentes segue a distribuição binomial negativa, a distribuição *a priori* conjugada para os parâmetros locais θ_m é a distribuição beta, para esta aplicação assume-se que os hiperparâmetros da distribuição beta são $c_0 = c_1 = 1$.

Para o modelo MPBBn as probabilidade de corte das covariáveis são apresentadas na

Tabela 4.5: Probabilidade *a posteriori* para as partições da covariável x_3 considerando os dados de melanoma para o modelo MPBPoi.

Partições	Probabilidade *a posteriori*	
	Cadeia 1	Cadeia 2
{1,2,3,4}	0,000	0,000
{1},{2,3,4}	0,002	0,002
{1,2},{3,4}	0,164	0,163
{1,2,3},{4}	0,766	0,766
{1},{2},{3,4}	0,002	0,002
{1},{2,3},{4}	0,033	0,035
{1,2},{3},{4}	0,032	0,032
{1},{2},{3},{4}	0,001	0,002

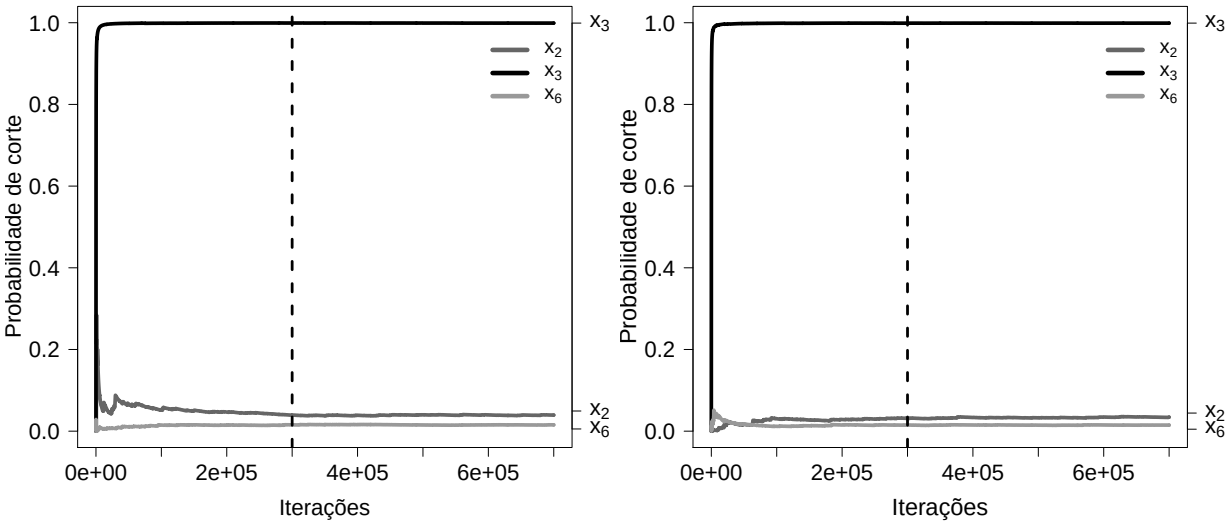
Tabela 4.6. Similarmente aos modelos de MPBBi e MPBPoi observa-se que a variável x_3 é dividida pela tesselação por hiperplanos, na maiorias das vezes, o que mostra que x_3 é uma covariável que tem efeito na fração de cura o que se traduz na probabilidade de corte *a posteriori* dessa variável ser 1 ou próxima de 1.

Observa-se que, independentemente dos valores assumidos para o parâmetro τ tem-se que a probabilidade de corte de x_3 permanece constante . Porém a probabilidade de corte de x_2 muda de acordo com os valores assumidos de τ e, neste caso, à medida que τ cresce a probabilidade de corte de x_2 diminui.

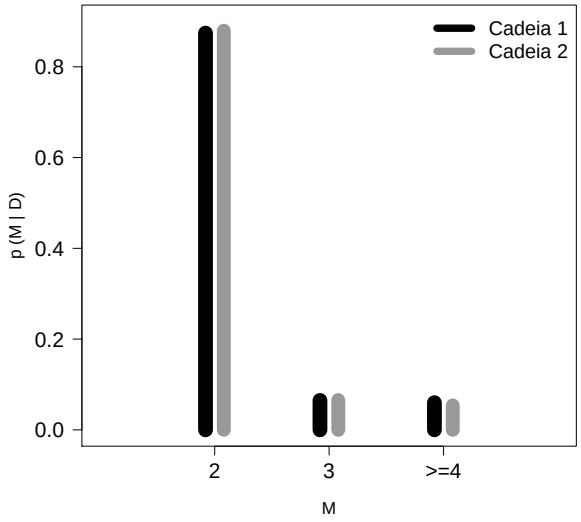
Tabela 4.6: Probabilidade de corte das covariáveis do conjunto de dados de melanoma seguindo o modelo MPBBn.

	Variáveis					
τ	x_1	x_2	x_3	x_4	x_5	x_6
1	0,020	0,256	1,000	0,019	0,017	0,092
3	0,007	0,256	1,000	0,007	0,009	0,035
7	0,002	0,146	1,000	0,002	0,004	0,016
13	0,001	0,078	0,999	0,001	0,001	0,006

A Tabela 4.7 mostra as probabilidades *a posteriori* para cada uma das partições da covariável x_3. O agrupamento que tem maior probabilidade *a posteriori* é a partição composta pelos grupos $(\{1,2,3\},\{4\})$. Conforme o valor de τ cresce a probabilidade *a*



(a) (b)



(c)

Figura 4.2: Evolução da probabilidade corte na (a) cadeia 1 e (b) cadeia 2. (c) Probabilidade *a posteriori* do número de regiões, para os dados de melanoma para o modelo MPBPoi.

posteriori do agrupamento $(\{1,2,3\},\{4\})$ também cresce. Uma situação inversa acontece com o agrupamento $(\{1,2\},\{3,4\})$.

A Tabela 4.8 apresenta o critério LPML para os valores assumidos do parâmetro τ. De acordo com o critério LPML pode-se observar que, se τ assume valores maiores que 1, o ajuste dos modelos aos dados não melhora em relação a $\tau = 1$ e desta forma o modelo MPBBn com $\tau = 1$ tem um ajuste melhor que os restantes modelos de MPBBn com valores

Tabela 4.7: Probabilidade *a posteriori* para as partições da covariável x_3 considerando os dados de melanoma para o modelo MPBBn.

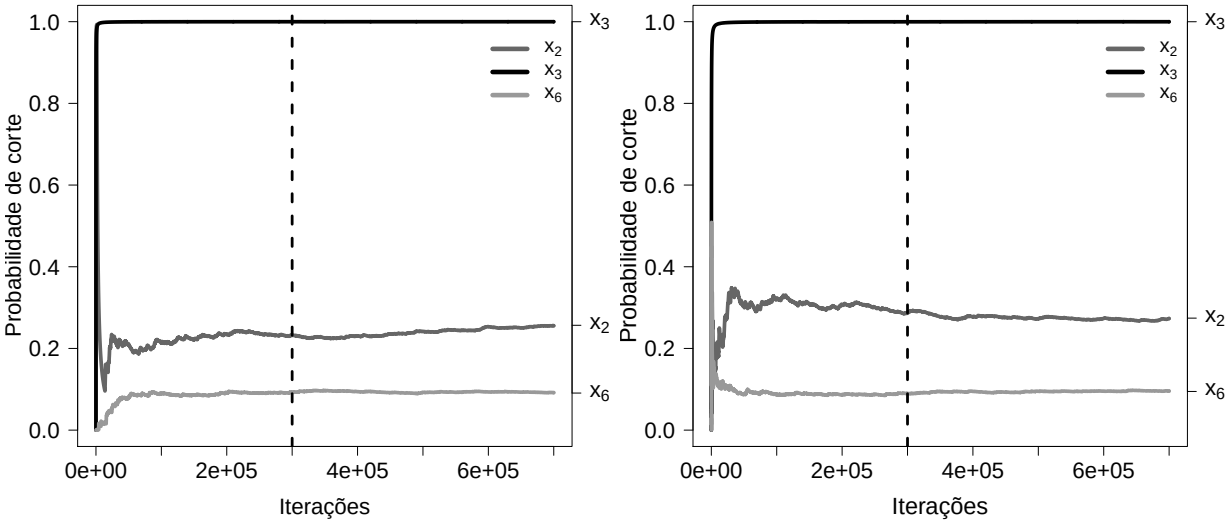
Partições	Probabilidade a *posteriori*			
	$\tau = 1$	$\tau = 3$	$\tau = 7$	$\tau = 13$
{1,2,3,4}	0,000	0,000	0,000	0,000
{1},{2,3,4}	0,002	0,001	0,001	0,001
{1,2},{3,4}	0,214	0,156	0,147	0,127
{1,2,3},{4}	0,340	0,644	0,771	0,827
{1},{2},{3,4}	0,028	0,006	0,001	0,001
{1},{2,3},{4}	0,191	0,085	0,034	0,019
{1,2},{3},{4}	0,170	0,098	0,044	0,024
{1},{2},{3},{4}	0,056	0,010	0,002	0,001

maiores que 1. Observa-se que se $\tau = 1$ em (4.6) obtém-se a distribuição geométrica.

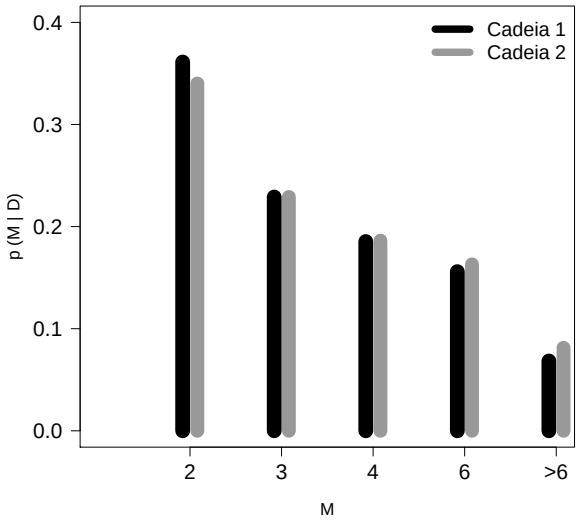
Tabela 4.8: Critério LPML para os modelos MPBBn.

	$\tau = 1$	$\tau = 3$	$\tau = 7$	$\tau = 13$
LPML	-519,892	-521,086	-521,285	-521,513

Os gráficos mostrados na Figura 4.3 são feitos considerando-se o modelo MPBGeo. A Figura 4.3(a) e 4.3(b) apresentam a evolução da probabilidade de corte das covariáveis x_2, x_3 e x_6 ao longo da simulação MCMC para cadeia 1 e cadeia 2, respectivamente. A Figura 4.3(c) mostra a probabilidade *a posteriori* do número de regiões na tesselação. Também o número de regiões M com maior probabilidade é quando $M = 2$, uma situação similar se apresentou nos modelos MPBBi e MPBPoi.



(a) (b)



(c)

Figura 4.3: Evolução da probabilidade corte na (a) cadeia 1 e (b) cadeia 2. (c) Probabilidade *a posteriori* do número de regiões, para os dados de melanoma para o modelo MPBGeo.

Resultados para o modelo MPBLg

Supondo-se que N segue uma distribuição logarítmica, considerou-se uma distribuição beta com parâmetros $d_0 = d_1 = 1$ como distribuição *a priori* para os parâmetros locais. As probabilidades de corte das covariáveis para o modelo MPBLg são apresentadas na Tabela 4.9. Observa-se que os hiperplanos dividem x_3, na maioria das vezes, na simulação MCMC pelo fato de a probabilidade de corte ser 1. Nesse caso, a variável x_3 tem um efeito significativo no modelo. Nesse sentido, as variáveis x_2 e x_6 fornecem pouca informação no modelo devido ao fato de que suas probabilidade de corte são relativamente baixas (0,134 e 0,128 respectivamente). As variáveis restantes têm probabilidades de corte próximas de zero, portanto essas variáveis não são informativas para modelo MPBLg.

Tabela 4.9: Probabilidade de corte para cada uma das covariáveis no modelo MPBLg.

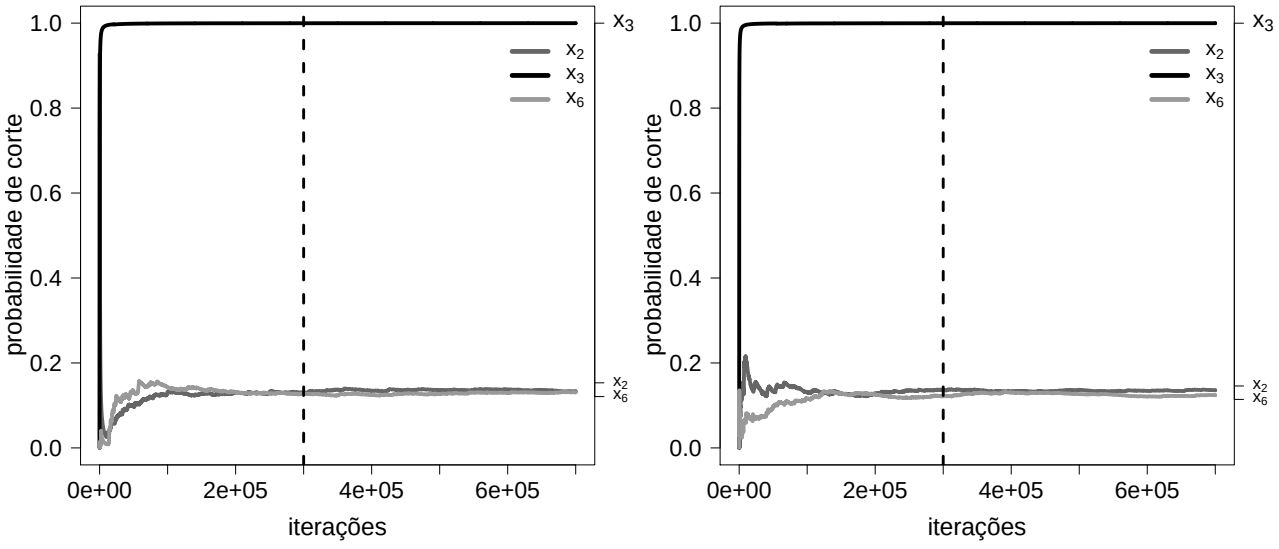
	Variáveis					
	x_1	x_2	x_3	x_4	x_5	x_6
Cadeia 1	0,010	0,133	1,000	0,012	0,009	0,131
Cadeia 2	0,011	0,136	1,000	0,011	0,009	0,124
Média	0,010	0,134	1,000	0,012	0,009	0,128

As probabilidades *a posteriori* para cada partição de x_3 são apresentadas na Tabela 4.10. É importante observar que, nos modelos MPBBi , MPBPoi e MPBBn, a partição para x_3 composta pelos subconjuntos $(\{1,2,3\},\{4\})$ tem uma probabilidade *a posteriori* maior em relação as outras partições. Porém é interessante observar que no modelo MPBLg a partição formada pelos subconjuntos $(\{1,2\},\{3,4\})$ tem maior probabilidade que a partição $(\{1,2,3\},\{4\})$. Outras partições que se destacam no modelo MPBLg são $(\{1\},\{2,3\},\{4\})$ e $(\{1\},\{2\},\{3,4\})$ e pode-se notar também, que o modelo de partição bayesiana identifica um ponto de mudança nos grupos em torno da categoria 2 para o modelo MPBLg, sendo que esta característica não foi identificada nos outros modelos.

Tabela 4.10: Probabilidade *a posteriori* para as partições da covariável x_3 considerando os dados de melanoma para o modelo MPBLg.

Partições	Probabilidade *a posteriori*	
	Cadeia 1	Cadeia 2
{1,2,3,4}	0,000	0,000
{1},{2,3,4}	0,016	0,011
{1,2},{3,4}	0,486	0,487
{1,2,3},{4}	0,090	0,095
{1},{2},{3,4}	0,133	0,130
{1},{2,3},{4}	0,154	0,154
{1,2},{3},{4}	0,074	0,079
{1},{2},{3},{4}	0,047	0,044

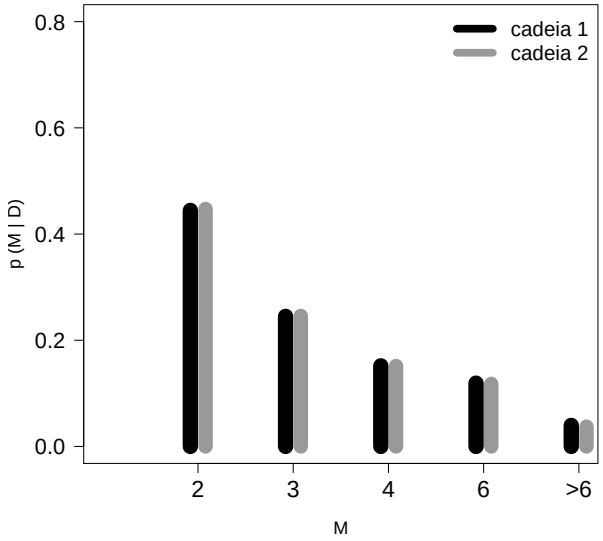
A Figura 4.4(a) e 4.4(b) apresentam a evolução da probabilidade de corte das covariáveis x_2, x_3 e x_6 ao longo da simulação MCMC para a cadeia 1 e cadeia 2, respectivamente.



(a) (b)

Figura 4.4: Evolução da probabilidade corte na (a) cadeia 1 e (b) cadeia 2.

A Figura 4.5 mostra a probabilidade *a posteriori* do número M de regiões na tesselação. A maior probabilidade *a posteriori* para M corresponde quando o número de regiões na tesselação é dois, $M = 2$.



(a)

Figura 4.5: Probabilidade *a posteriori* do número de regiões, para os dados de melanoma para o modelo MPBLg.

Comparação dos modelos

Observou-se que, de acordo com as Tabelas 4.2, 4.5, 4.7 e 4.10 o modelo MPB fornece uma interpretação em relação à composição da partição da variável categoria do nódulo (x_3). De acordo com os modelos MPBBi, MPBPoi e MPBBn, os pacientes que estão no estagio 1,2 e 3 compõem um grupo homogêneo, com uma probabilidade maior, em contraste com as outras partições de x_3 (veja Tabela 3.2).

Neste sentido, intuitivamente pode-se interpretar que a probabilidade de cura para os indivíduos no estágio 1,2 e 3 é a mesma, no entanto o modelo de partição bayesiana indica que os pacientes no estágio 4 têm um comportamento diferente. Para o modelo MPBLg a interpretação é similar, porém considerando-se o agrupamento $(\{1, 2\}, \{3, 4\})$.

Na Tabela 4.11 são apresentados os resumos *a posteriori* dos parâmetros α e λ da distribuição Weibull, tais como a média, desvio padrão (DP), o intervalo de maior densidade *a posteriori* de 95% (95% HPD) e o critério LPML para os modelos MPBBi com $K = 10$,

MPBPoi, MPBGeo e o modelo MPBLg. O critério LPML fornece evidência a favor do modelo MPBLg e, como segundo modelo, tem-se o modelo MPBGeo.

Tabela 4.11: Resumos das distribuições *a posteriori* dos parâmetros da distribuição Weibull para o conjunto de dados de melanoma.

Modelo	LPML	Parâmetro	Média	DP	95%HPD
MPBBi†	-521,775	α	1,599	0,109	$(1,394; 1,820)$
		λ	-1,295	0,125	$(-1,532; -1,050)$
MPBPoi	-521,482	α	1,721	0,116	$(1,495; 1,947)$
		λ	-1,645	0,135	$(-1,920; -1,388)$
MPBGeo	-519,892	α	1,869	0,125	$(1,624; 2,105)$
		λ	-2,069	0,125	$(-2,390; -1,757)$
MPBLg	-519,004	α	2,040	0,136	$(1,766; 2,293)$
		λ	-2,454	0,213	$(-2,890; -2,071)$

$^{\dagger}K = 10$

A Figura 4.6 mostra as estimativas de K-M da função de sobrevivência, assim como a estimativa obtida dos modelos MPBBi com $K = 10$, MPBPoi e MPBGeo para a covariável x_3 considerando o agrupamento $\{1, 2, 3\}$ e $\{4\}$ e na Figura 4.6(d) a estimativa da função de sobrevivência do modelo MPBLg porém levando em conta o agrupamento $\{1, 2\}$ e $\{3, 4\}$ da variável x_3.

A Tabela 4.12 apresenta as estimativas da fração de cura para os modelos MPBBer, MPBPoi, MPBGeo considerando a covariável categoria do nódulo (x_3) e o agrupamento $\{\{1, 2, 3\}, \{4\}\}$. No caso do modelo MPBLg a estimativa da fração de cura é calculada levando em conta o agrupamento $\{\{1, 2\}, \{3, 4\}\}$.

Tabela 4.12: Estimativa da fração de cura para o conjunto de dados de melanoma.

Modelo	Categoria do nodulo (x_3)	
	{1,2,3}	{4}
MPBBer	0,566	0,307
MPBPoi	0,559	0,281
MPBGeo	0,549	0,305
	{1,2}	{3,4}
MPBLg	0,583	0,401

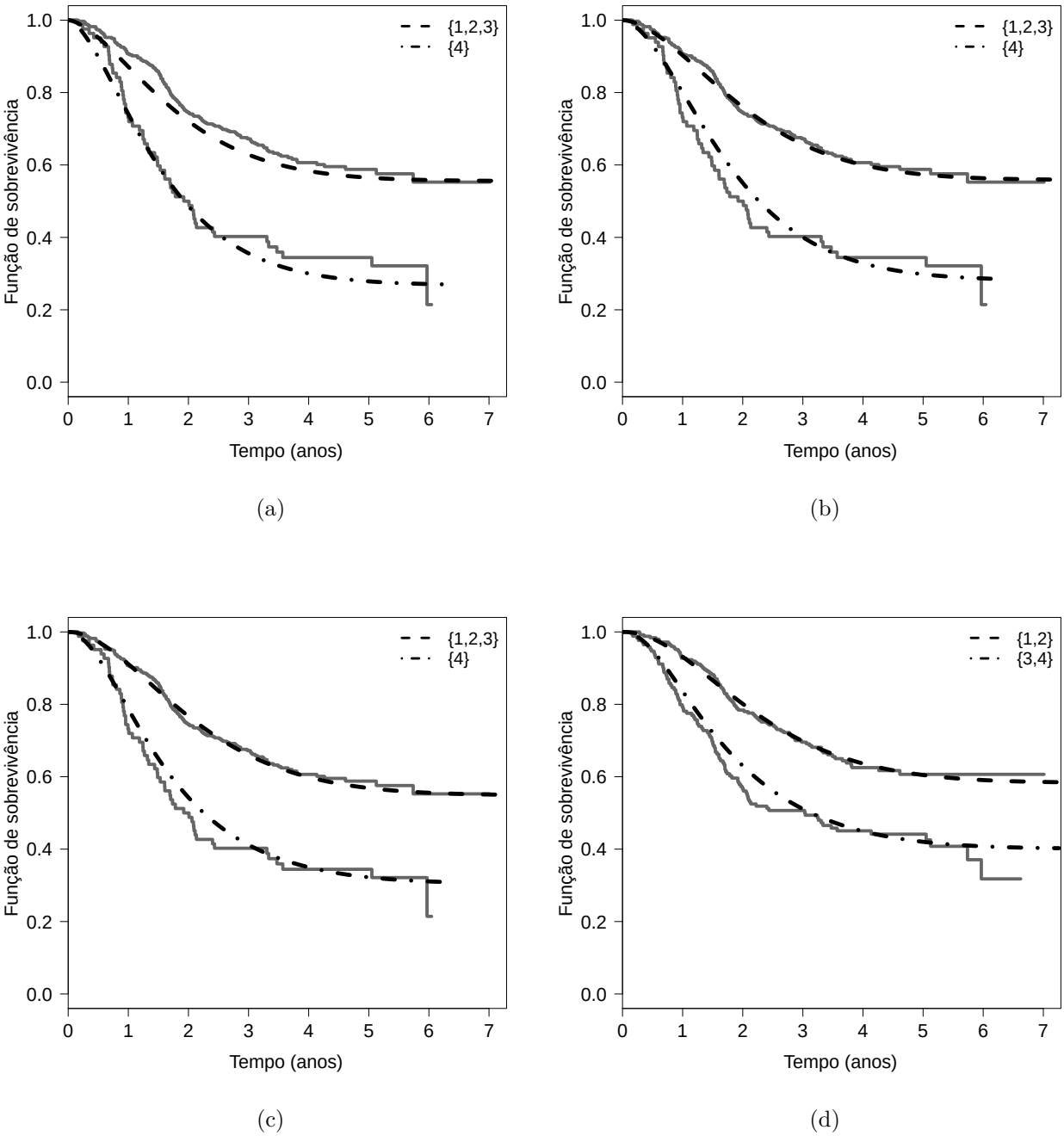


Figura 4.6: Curvas de K-M estratificado de acordo com a covariável x_3 para o agrupamento $\{1, 2, 3\}$ e $\{4\}$: (a) modelo MPBBi com $K = 10$ (b) modelo MPBPoi e (c) modelo MPBGeo. Em (d) mostra a estimativa da função de sobrevivência seguindo o modelo MPBLg considerando o agrupamento $\{1, 2\}$ e $\{3, 4\}$.

4.4.2 Dados de leucemia

Nesta seção, foi aplicado o modelo de partição bayesiana para os dados de leucemia que foram analisados na Seção 2.4. Porém assume-se que a covariável ano de transplante

de médula óssea são *a priori* grupos separados e deixou-se que o modelo MPB indique os grupos que são homogêneos, de acordo com sua probabilidade *a posteriori* .

Para obter as estimativas bayesianas dos parâmetros da distribuição Weibull foram adotadas as distribuições *a priori* como na Seção (4.1.1), assim $\alpha \sim \text{Ga}(0,1,0,1)$ e $\lambda \sim \text{N}(0,100)$. Para o número de regiões M na tesselação, foi assumida uma distribuição geométrica com média 10, $M \sim \text{Geo}(0.1)$.

Foram geradas duas cadeias independentes com 700000 iterações para os modelos de longa duração com partição bayesiana propostos na Seção 4.2. As primeiras 200000 foram descartadas como iterações *burn-in*, e foi adotado um salto de tamanho 100, conduzindo a uma amostra final de tamanho 5000 para cada cadeia de cada caso. No começo do algoritmo consideramos, $\boldsymbol{N} = (1,\ldots,1)$ e número de regiões $M = 1$.

Para monitorar a convergência foram consideradas as probabilidades de corte em ambas as cadeias, assim como as probabilidades *a posteriori* dos agrupamentos da variável x_1. A convergência dos parâmetros da distribuição Weibull foi monitorada com o auxilio do fator de redução de escala ($\widehat{R}$) proposto por Gelman & Rubin (1992).

Resultados para o modelo MPBBi

No caso que o número de causas latentes N segue a distribuição binomial foi adotada a distribuição beta com parâmetros $a_0 = a_1 = 1$ como distribuição *a priori* para θ_m. Foi assumido que parâmetro K da distribuição binomial é fixo e que o conjunto de valores para K é dada por $\{1,5,15,30\}$.

A Tabela 4.13 apresenta a probabilidade de corte de cada uma das covariáveis para diferentes valores de K para o modelo MPBBi. Observa-se que as covariáveis x_1 (ano de transplante de médula ossea) e x_2 (idade do paciente) tem um efeito significativo na fração de cura quando $K = 1$ e portanto as probabilidades *a posteriori* dessas variáveis são próximas de 1. Para valores de $K > 1$ a probabilidade de corte para x_1 permanece constante(próximo de 1), porém a probabilidade corte de x_2 é menor em relação a x_1 cada vez que o valor de K cresce. Neste cenário, pode-se interpretar que x_1 tem um efeito na fração de cura para o modelo MPBBi independentemente do valor assumido de K, embora um efeito inverso ocorra com x_2.

A variável x_1 é uma variável qualitativa ordinal com mais de duas categorias e, desta forma, a partição de x_1 pode ser feita considerando-se a ordem dos níveis. Porém consideramos x_1 como uma variável qualitativa nominal o que nos leva a ter no máximo $n_\rho = 5$ partições diferentes para x_1.

Tabela 4.13: Probabilidade de corte para cada covariável no modelo MPBBi para o conjunto de dados de leucemia.

	Variáveis			
K	x_1	x_2	x_3	x_4
1	0,998	0,853	0,004	0,004
5	0,997	0,401	0,003	0,001
15	0,999	0,077	0,000	0,000
30	0,998	0,022	0,000	0,000

A Tabela 4.14 apresenta as probabilidades *a posteriori* para os agrupamentos de x_1. Nesse caso, a partição composta por $\{1\}, \{2, 3\}$ tem a maior probabilidade *a posteriori* em relação às outras. Para o modelo MPBBi existe diferença entre os indivíduos que estão no nível $\{1\}$ e o grupo formado por $\{2, 3\}$.

Tabela 4.14: Probabilidade *a posteriori* para os agrupamentos da variável x_1 no modelo MPBBi para os dados de leucemia.

Partições	Probabilidade *a posteriori*			
	$K = 1$	$K = 5$	$K = 15$	$K = 30$
$\{1, 2, 3\}$	0,002	0.004	0.001	0.002
$\{1\}, \{2, 3\}$	0,977	0,987	0,994	0,996
$\{1, 2\}, \{3\}$	0,000	0,000	0,000	0,000
$\{1, 3\}, \{2\}$	0,000	0,000	0,000	0,000
$\{1\},\{2\},\{3\}$	0,0205	0,009	0,004	0,002

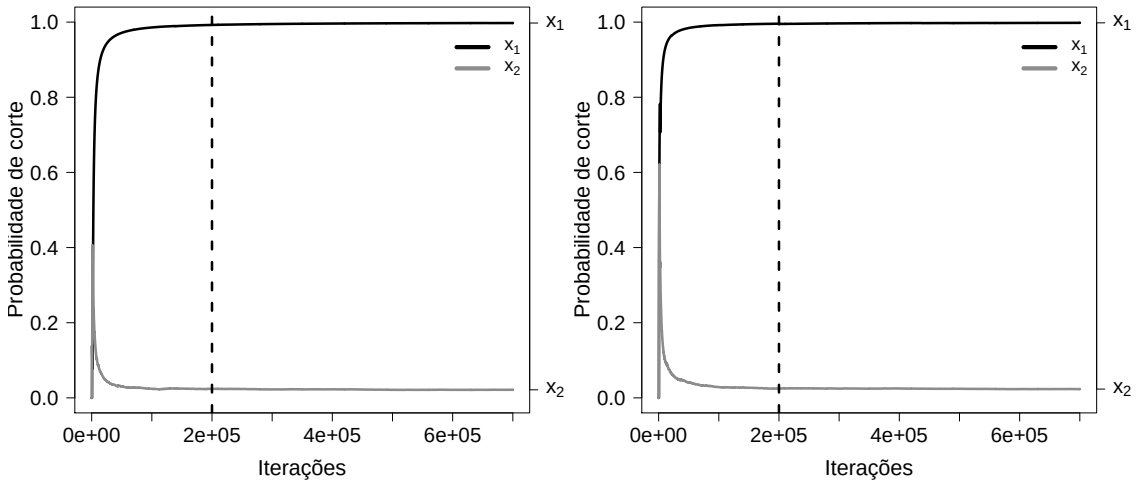
A Tabela 4.15 apresenta a estatística LPML para o modelo MPBBi para os diferentes valores fixados. Foi observado que, à medida que o valor de K aumenta, o ajuste também melhora, porém o valor da estatística LPML tende a se estabilizar. Por isso, escolhemos o modelo MPBBi com $K = 30$ sendo o modelo que se ajusta melhor aos dados.

Tabela 4.15: Critério LPML para os modelos MPBBi para os dados de leucemia.

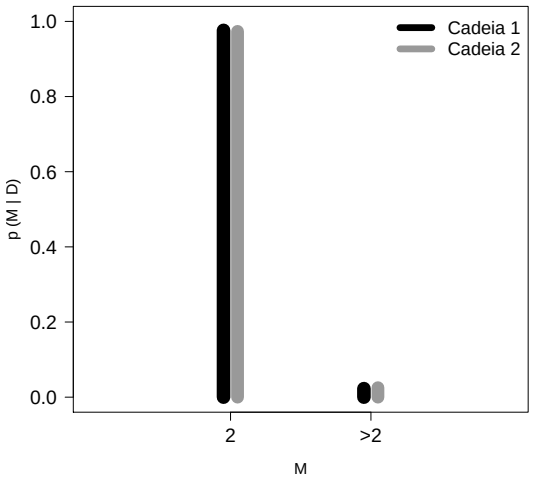
	$K = 1$	$K = 5$	$K = 15$	$K = 30$
LPML	-1757,346	-1754,302	-1753,838	-1753,492

A Figura 4.7(a) e 4.7(b) mostra a evolução da probabilidade de corte das covariáveis x_1 e x_2 ao longo da simulação MCMC para cadeia 1 e cadeia 2, respectivamente. A Figura

4.7(c) mostra a probabilidade *a posteriori* do número M de regiões na tesselação. É importante observar que a maior probabilidade *a posteriori* para M corresponde, quando o número de regiões na tesselação é dois, $M = 2$. Os gráficos apresentados na Figura 4.7 foram feitos considerando-se o modelo MPBBi com $K = 30$.



(a) (b)



(c)

Figura 4.7: Evolução da probabilidade corte na (a) cadeia 1 e (b) cadeia 2. (c) Probabilidade *a posteriori* do número de regiões, para os dados de melanoma para o modelo MPBBi com $K = 30$

Resultados para o modelo MPBPoi

Se o número de causas latentes segue a distribuição Poisson, considera-se uma distribuição gama com hiperparâmetros $b_0 = b_1 = 1$ como distribuição *a priori* para os parâmetros locais.

As probabilidades de corte para o modelo MPBPoi são apresentadas na Tabela 4.16. Os hiperplanos dividem as variáveis x_1 e x_2, na maioria das vezes, na simulação MCMC, notando-se que a probabilidade de corte da variável x_1 é próximo de 1 e, para x_2 a probabilidade de corte é dada por $0,886$.

Tabela 4.16: Probabilidade de corte para as variáveis preditoras no modelo MPBPoi considerando os dados de leucemia.

	Variáveis			
	x_1	x_2	x_3	x_4
Cadeia 1	1,000	0,887	0,003	0,002
Cadeia 2	0,999	0,884	0,003	0,002
Média	1,000	0,886	0,003	0,002

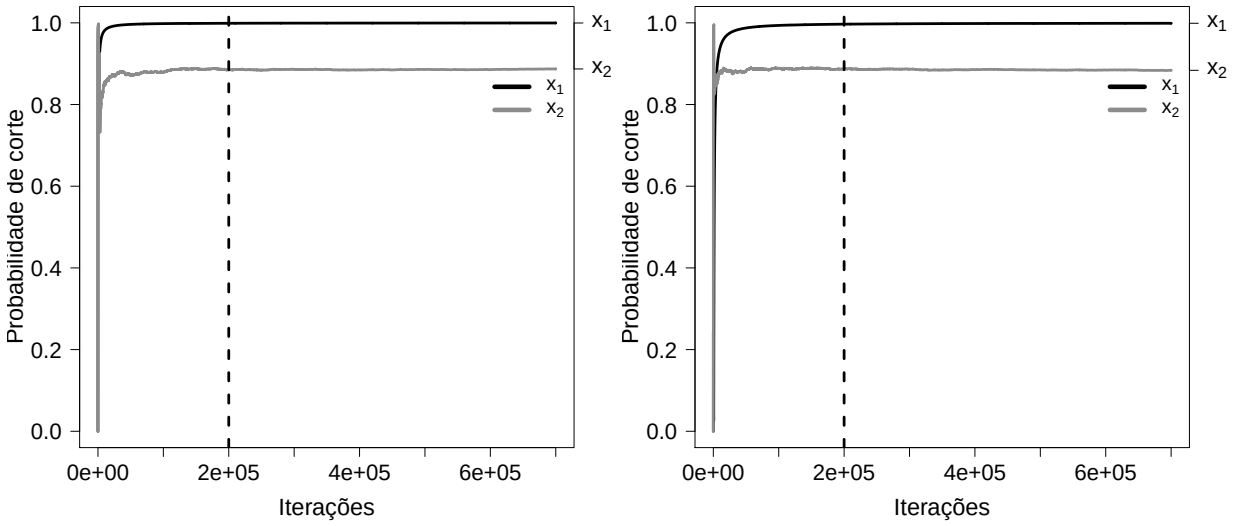
As probabilidades *a posteriori* dos agrupamentos, para a variável x_1 são apresentadas na Tabela 4.17. A partição com maior probabilidade *a posteriori* é formada pelos grupos $\{1\}$ e $\{2,3\}$. Desta forma, o modelo MPBPoi identifica esse agrupamento como o mais plausível para o conjunto dos dados.

Tabela 4.17: Probabilidade *a posteriori* para os agrupamentos da veriável x_1 no modelo MPBPoi para os dados de leucemia.

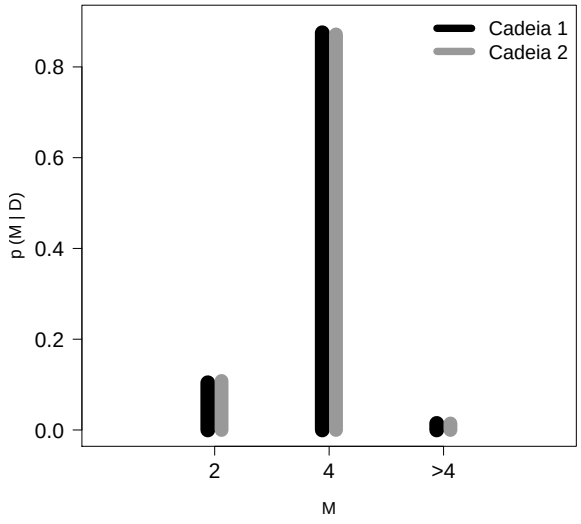
Partições	Probabilidade a *posteriori*		
	Cadeia 1	Cadeia 2	Média
$\{1,2,3\}$	0,000	0,001	0,000
$\{1\},\{2,3\}$	0,981	0,980	0,980
$\{1,2\},\{3\}$	0,000	0,000	0,000
$\{1,3\},\{2\}$	0,000	0,000	0,000
$\{1\},\{2\},\{3\}$	0,018	0,018	0,018

A Figura 4.8(a) e 4.8(b) mostra a evolução da probabilidade de corte das covariáveis x_1 e x_2 ao longo da simulação MCMC para cadeia 1 e cadeia 2, respectivamente. A Figura 4.8(c) mostra a probabilidade *a posteriori* do número M de regiões na tesselação. Nota-se

que a maior probabilidade *a posteriori* para M corresponde quando o número de regiões na tesselação é quatro, $M = 4$.



(a) (b)



(c)

Figura 4.8: Evolução da probabilidade corte na (a) cadeia 1 e (b) cadeia 2. (c) Probabilidade *a posteriori* do número de regiões, para os dados de melanoma para o modelo MPBPoi

Resultados para o modelo MPBBn

Caso o número de riscos latentes siga a distribuição binomial negativa, a distribuição *a priori* conjugada para os parâmetros locais θ_m é a distribuição beta, em que se assume que os parâmetros da distribuição beta são dadas por $c_0 = c_1 = 1$.

As probabilidades de corte para variáveis do conjunto de leucemia, considerando o modelo MPBBn, são apresentadas na Tabela 4.18. Nota-se que independentemente do valor assumido para o parâmetro τ da distribuição binomial negativa, a variável x_1 tem um efeito no modelo. Em seguida, observa-se que a covariável x_2 tem influência no modelo quando $\tau = 1$ porém, à medida que τ cresce, tem-se que a probabilidade de corte de x_2 decresce, assim a influência da variável x_2 no modelo depende do valor de τ.

Tabela 4.18: Probabilidade de corte para cada covariável para o modelo MPBBn para o conjunto de dados de leucemia.

	Variáveis			
τ	x_1	x_2	x_3	x_4
1	1,000	0,835	0,003	0,002
7	1,000	0,252	0,001	0,000
13	0,998	0,096	0,000	0,000
30	0,998	0,022	0,000	0,000

Para o modelo MPBBn a partição $\{1\}, \{2, 3\}$ da variável x_1 tem a maior probabilidade entre as outras partições como pode ser visto na Tabela 4.19. É possível afirmar que o modelo MPBBn indica que a taxa de cura para pacientes no nível 1 é diferente daquela dos pacientes dos níveis $\{2, 3\}$ e esta característica não é alterada para os diferentes valores assumidos do parâmetro τ.

A Tabela 4.20 apresenta a estatística LPML para o modelo MPBBn para os diferentes valores fixados. Nota-se que para valores de $\tau > 1$, o ajuste do modelo MPBBn não melhora. Razão pela qual foi escolhido o modelo MPBBn com $\tau = 1$, sendo o melhor modelo que se ajusta aos dados.

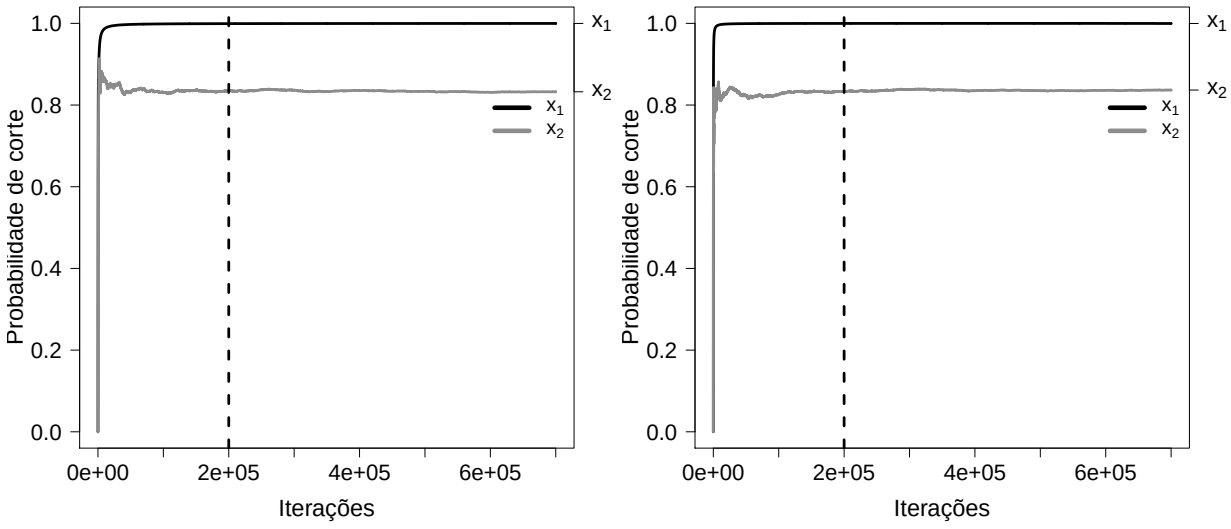
Tabela 4.19: Probabilidade *a posteriori* para os agrupamentos da veriável x_1 no modelo MPBBn para os dados de leucemia.

Partições	Probabilidade a *posteriori*			
	$\tau = 1$	$\tau = 7$	$\tau = 13$	$\tau = 30$
$\{1, 2, 3\}$	0,000	0,000	0,002	0,002
$\{1\}, \{2, 3\}$	0,980	0,990	0,994	0,995
$\{1, 2\}, \{3\}$	0,000	0,000	0,000	0,000
$\{1, 3\}, \{2\}$	0,000	0,000	0,000	0,000
$\{1\}, \{2\}, \{3\}$	0,019	0,008	0,005	0,002

Tabela 4.20: Critério LPML para os modelos MPBBn para os dados de leucemia

	$\tau = 1$	$\tau = 7$	$\tau = 13$	$\tau = 30$
LPML	-1743,089	-1752,09	-1752,666	-1752,87

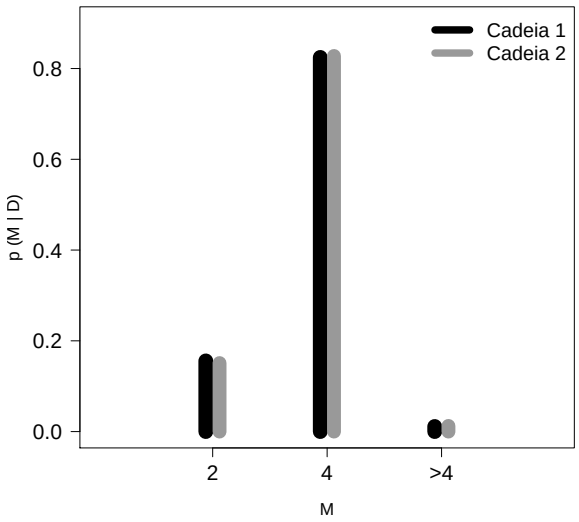
A Figura 4.9(a) e 4.9(b) mostra a evolução da probabilidade de corte das covariáveis x_1 e x_2 ao longo da simulação MCMC para cadeia 1 e cadeia 2, respectivamente.



(a) (b)

Figura 4.9: Evolução da probabilidade corte na (a) cadeia 1 e (b) cadeia 2.

A Figura 4.10 mostra a probabilidade *a posteriori* do número M de regiões na tesselação. Foi observado que a maior probabilidade *a posteriori* para M corresponde quando o número de regiões na tesselação é quatro, $M = 4$.



(a)

Figura 4.10: Probabilidade *a posteriori* do número de regiões, para os dados de leucemia para o modelo MPBGeo.

Resultados para o modelo MPBLg

Se N segue uma distribuição logarítmica, considera-se uma distribuição beta com parâmetros $d_0 = d_1 = 1$ sendo a distribuição *a priori* para os parâmetros locais.

Para o modelo MPBLg, as probabilidades de corte para as covariáveis são apresentadas na Tabela 4.21. Observou-se que as variáveis x_1 e x_2 têm um efeito na fração de cura, porém a probabilidade *a posteriori* de x_1 (próximo a 1) é maior do que a da variável x_2 e, o que intuitivamente, indica que, para o modelo MPBLg a variável x_1 tem um maior efeito sobre a fração de cura. As variáveis x_1 e x_2 também foram selecionadas pelos modelos de partição MPBBi, MPBPoi, MPBBn, assim, independentemente do modelo adotado, há evidência de que x_1 e x_2 são variáveis selecionadas para modelar a taxa de cura.

Tabela 4.21: Probabilidade de corte para cada covariável para o modelo MPBLg para o conjunto de dados de leucemia..

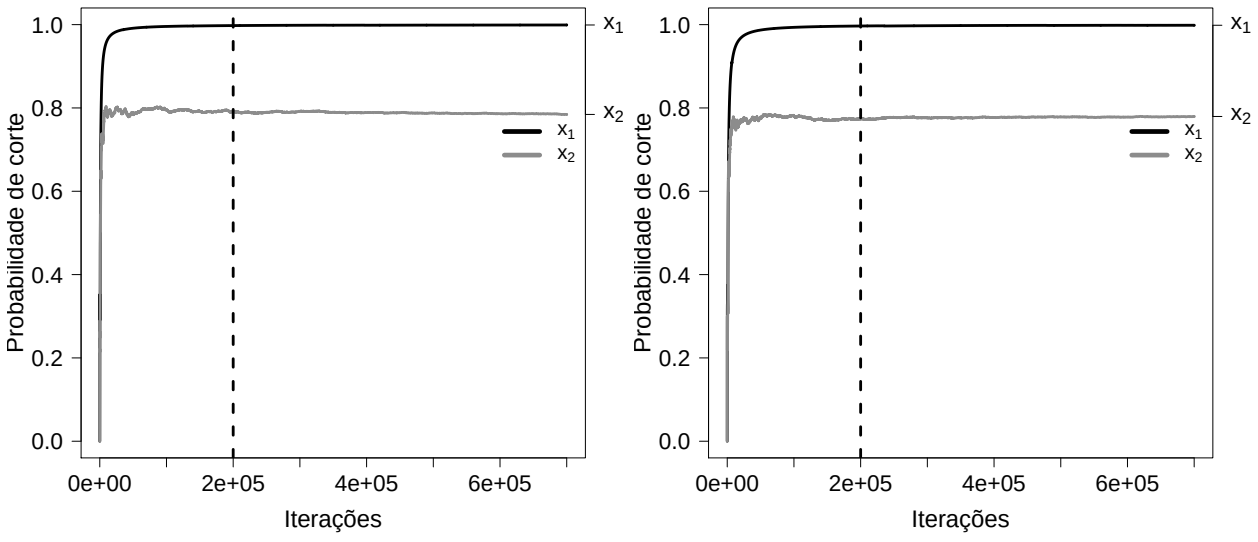
	Variáveis			
	x_1	x_2	x_3	x_4
Cadeia 1	0,999	0,784	0,004	0,003
Cadeia 2	0.999	0,780	0,004	0,003
Média	0,999	0,782	0,004	0,003

A Tabela 4.22 mostra as probabilidades *a posteriori* dos grupos da variável x_1. Como os outros modelos de partição apresentados anteriormente, para o modelo MPBLg a partição composta pelas categorias $\{1\}, \{2, 3\}$ tem a maior probabilidade *a posteriori* em relação às outras partições de x_1.

Tabela 4.22: Probabilidades *a posteriori* para os agrupamentos da variável x_1 para o modelo MPBLg para os dados de leucemia.

Partições	Probabilidade a *posteriori*		
	Cadeia 1	Cadeia 2	Média
{1,2,3}	0,001	0,002	0,002
{1},{2,3}	0,967	0,964	0,966
{1,2},{3}	0,000	0,000	0,000
{1,3},{2}	0,000	0,000	0,000
{1},{2},{3}	0,032	0,034	0,033

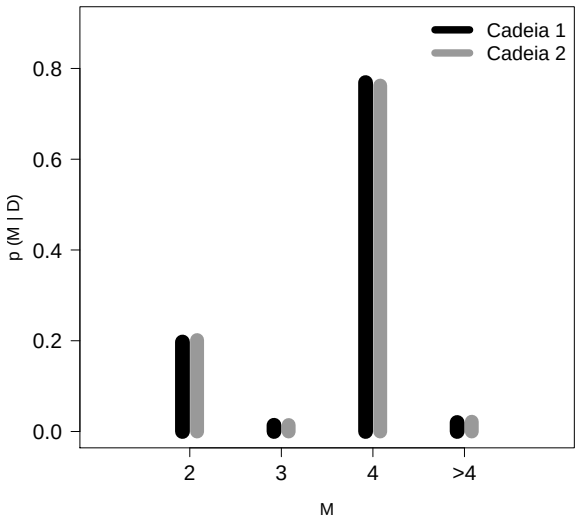
A Figura 4.11(a) e 4.11(b) apresentam a evolução da probabilidade de corte das covariáveis x_1 e x_2 ao longo da simulação MCMC para cadeia 1 e cadeia 2, respectivamente.



(a) (b)

Figura 4.11: (a) e (b) Mostram a evolução da probabilidade de corte das covariáveis para cadeia 1 e 2 respectivamente no modelo MPBLg

A Figura 4.12 mostra a probabilidade *a posteriori* do número de regiões na tesselação $\boldsymbol{\mathcal{T}}$ considerando as duas cadeias geradas pelo amostrador MCMC. Nota-se que o número de regiões com maior probabilidade *a posteriori* é quando $M = 4$.



(a)

Figura 4.12: Probabilidade *a posteriori* do número de regiões na tesselação para os dados de leucemia considerando o modelo MPBLg.

Comparação dos modelos

Os modelos MPBBi, MPBPoi, MPBBn, MPBLg confirmaram que as categorias {2} e {3} fazem parte de um mesmo grupo com uma alta probabilidade. Para os modelos de partição propostos, têm-se que, além de selecionar variáveis preditoras, também podem alocar os indivíduos em agrupamentos (no caso de variáveis qualitativas), em que esses indivíduos podem ser considerados homógenos em relação a uma característica que neste caso é taxa de cura.

Na Tabela 4.23 são apresentados os resumos *a posteriori* dos parâmetros α e λ da distribuição Weibull, tais como a média, desvio padrão (DP), o intervalo de maior densidade *a posteriori* de 95% (95% HPD) e o critério LPML para os modelos MPBBi com $K = 30$, MPBPoi, MPBGeo e o modelo MPBLg. De acordo com o critério LPML, o modelo MPBLg tem um melhor ajuste entre os modelos considerados e como segundo melhor modelo tem-se o modelo MPBGeo.

A Figura 4.13 apresenta as estimativas de K-M da função de sobrevivência e estimativa

Tabela 4.23: Resumos das distribuições *a posteriori* para os parâmetros da distribuição Weibull.

Modelo	LPML	Parâmetro	Média	DP	95% HPD
MPBBi†	-1753,492	α	0,815	0,024	$(0,768; 0,862)$
		λ	-0,139	0,048	$(-0,233 - 0,046)$
MPBPoi	-1748,865	α	0,818	0,024	$(0,773; 0,866)$
		λ	-0,149	0,048	$(-0,245; -0,056)$
MPBGeo	-1743,089	α	0,870	0,025	$(0,821; 0,917)$
		λ	-0,349	0,055	$(-0,459; -0,243)$
MPBLg	-1740,439	α	0,920	0,026	$(0,868; 0,971)$
		λ	-0,519	0,061	$(-0,643; -0,403)$

$^{\dagger}K = 30$

obtida do modelo MPBLg considerando a partição com maior probabilidade para a covariável x_1, isto é, a partição formada por $\{1\}$ e $\{2, 3\}$.

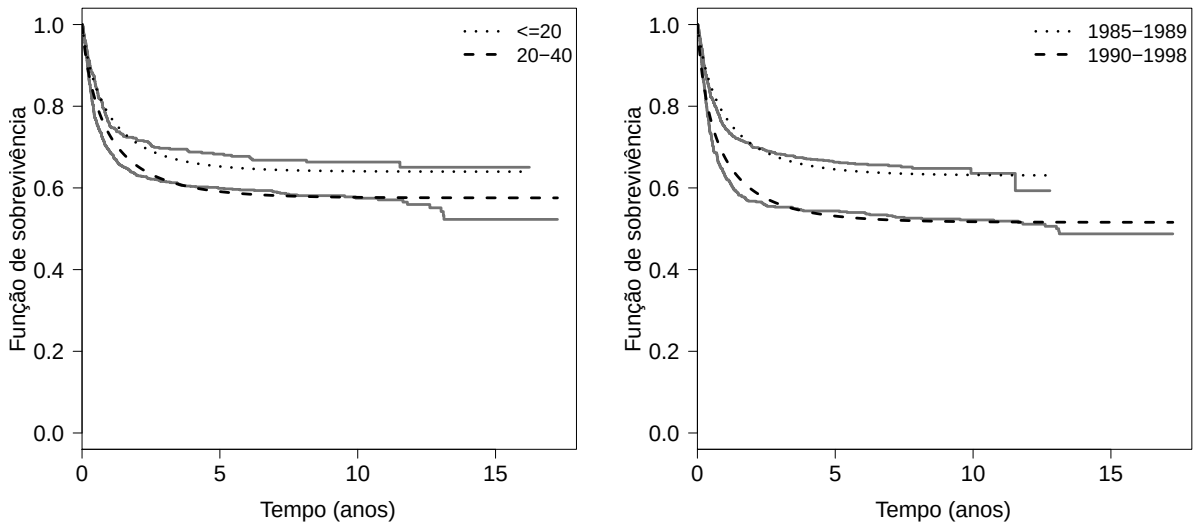


Figura 4.13: Estimativa de K-M da função de sobrevivência e estimativa da função de sobrevivência estratificado para as covariáveis idade (painel esuquerdo) e ano de transplante (painel direito) de acordo com o modelo MPBLg para os dados de pacientes com leucemia.

Observação 4.1. No Apêndice A e B são apresentados o histórico das cadeias e a densidades marginais *a posteriori* aproximadas para os parâmetros da distribuição Weibull para

os modelos ajustados, considerando-se conjunto de melanoma e leucemia, respectivamente.

Apresenta-se na Tabela 4.24 as estimativas da fração de cura. Assim para calcular a estimativa da fração de cura foram consideradas as variáveis x_1 e x_2, isto porque essas covariáveis tem um efeito maior na fração de cura. Nesse sentido, foi assumido que os pacientes que receberam o transplante de médula nos anos 1990-1994 e 1995-1998 são parte de um mesmo grupo isto pelo fato que a probabilidade que esses grupos sejam parte de um mesmo agrupamento é alta em todos os modelos ajustados.

Tabela 4.24: Estimativa da fração de cura para o conjunto de dados de leucemia.

Modelo	Ano de transplante (x_1)		Idade (x_2)	
	$1985-1989$	$1990-1998$	≤ 20	$20-40$
MPBBer	0,508	0,647	0,654	0,579
MPBPoi	0,498	0,641	0,651	0,566
MPBGeo	0,503	0,635	0,644	0,570
MPBLg	0,516	0,630	0,640	0,576

4.5 Comentários finais

Neste capítulo, foi proposta uma extensão local para o modelo de série de potências com fração de cura baseado no modelo de partição bayesiana, em que assumiu-se uma distribuição Weibull para os tempos de ocorrência para o evento de interesse. No modelo MPB, também foram consideradas covariáveis qualitativas. Nesse sentido foi proposta uma estratégia computacional para a simulação MCMC. A vantagem dese considerar a tesselação por hiperplanos no modelo partição bayesiana é que a seleção das covariáveis, que tem influência na fração de cura, é feita naturalmente dividindo-se as covariáveis que tem um efeito no modelo e exclui as variáveis que não têm um impacto significativo. A modelagem proposta foi aplicado a dois conjuntos de dados reais. Em ambos os conjuntos de dados, o modelo MPBLg se apresentou como o melhor modelo que se ajusta aos dados.

Capítulo 5

Considerações finais e propostas futuras de trabalho

5.1 Considerações finais

Neste trabalho, foram apresentados os resultados mais relevantes da teoria unificada de longa duração proposta por Tsodikov *et al.* (2003) e Rodrigues *et al.* (2009a). O modelo de mistura padrão e o modelo de risco acumulado limitado são casos particulares desta teoria.

Apresentou-se a metodologia do modelo de partição bayesiana proposto por Holmes *et al.* (1999, 2005), cuja característica especial é sua capacidade preditiva, como pode ser visto em Denison *et al.* (2002a), Hoggart & Griffin (2001), Hopcroft *et al.* (2009), entre outros.

Para obter uma partição no espaço preditor foi considerada uma tesselação por hiperplanos ortogonais ao vez da tesselação de Voronoi. Nesse sentido, a tesselação por hiperplanos conduz a seleção de covariáveis que influenciam a fração de cura nos modelos de longa duração propostos. Assim, se alguma covariável tem um efeito na fração de cura, um hiperplano divide essa covariavel e, desta forma, as variáveis preditoras são selecionadas no modelo. Além disso, o custo computacional para construir a tesselação por hiperplanos é menor em relação a tesselação de Voronoi.

Utilizando-se a metodologia de partição bayesiana, foi apresentada uma extensão do modelo proposto por Hoggart & Griffin (2001), por considerar variáveis qualitativas com mais de duas categorias e foram feitas modificações no amostrador MCMC, para lidar com esse tipo de variáveis. A metodologia proposta neste trabalho foi aplicada, a dois

conjuntos de dados reais.

As principais contribuições deste trabalho foram a extensão local dos modelos de longa duração baseada no modelo de partição bayesiana (Holmes *et al.*, 2005) em que se usou uma estrutura local no espaço preditor e se considerou uma família flexível para o número de causas latentes, a distribuição de série de potências . Além disso, foi desenvolvida uma estratégia computacional adequada para selecionar covariáveis qualitativas e quantitativas que são significativas no modelo. A programação da estratégia computacional, está baseada em linguagem de programação R (R Core Team, 2013).

5.2 Propostas futuras de trabalho

5.2.1 Dicotomização de uma variável continua no modelo de riscos proporcionais de Cox baseado no modelo de partição bayesiana

Um modelo usado amplamente em estudos clínicos e epidemiológicos, entre outros, é o modelo de riscos proporcionais de Cox (MRP). Assim, o modelo MRP relaciona o tempo até o evento de interesse de um individuo com um conjunto de covariáveis sendo definida por

$$h(t|\boldsymbol{x}) = h_0(t)\exp(\boldsymbol{x}^\top\boldsymbol{\beta}), \tag{5.1}$$

em que $h_0(t)$ é a função de risco base, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ é o vetor de coeficientes de regressão que descrevem os efeitos das covariáveis. Note-se que as covariáveis têm um efeito multiplicativo na função de risco. Logo, a função de sobrevivência $S(t|\boldsymbol{x})$ é dada por

$$S(t|\boldsymbol{x}) = \exp\left(-H_0(t)\exp(\boldsymbol{x}^\top\boldsymbol{\beta})\right),$$

em que $H_0(t)$ é a função de risco acumulado.

Considerando covariáveis discretas ou categóricas, pode-se interpretar facilmente o efeito dessas covariáveis no modelo dado em (5.1). Por exemplo, seja x uma variável preditora binária que representa se um indivíduo é diabético ou não. Neste caso pode-se avaliar e interpretar o efeito dessa covariável no tempo de sobrevivência do individuo considerando o modelo de Cox. Porém, no caso em que x é de natureza contínua a interpretação do modelo MRP é difícil e, neste cenário, usualmente os pesquisadores

discretizam x em duas ou mais categorias. No entanto, nosso interesse é dicotomizar a variável contínua x.

Existem várias propostas para dicotomizar variáveis contínuas para o modelo de Cox, como pode ser visto em Contal & O'Quigley (1999), Jensen & Lütkebohmert (2008) no contexto frequentista e Chen *et al.* (2014), sob perspectiva bayesiana.

Assim, para dicotomizar uma covariável contínua no modelo MRP foi usado o modelo MPB. Neste sentido, pretende-se olhar o modelo MRP de uma perspectiva local, assim, supõe-se que a tesselação $\boldsymbol{\mathcal{T}}$ divide o espaço preditor $\boldsymbol{\mathcal{X}}$ em M regiões e, de acordo com o modelo MPB, é necessário definir quais são os parâmetros locais em cada região da tesselação. Assim, seja n_m o número de indivíduos na região R_m e θ_m o parâmetro local, a função de risco para um individuo na região R_m é dada por

$$h(t_{mj}|\boldsymbol{x}_{mj}) = h_0(t_{mj})\theta_m, \quad \boldsymbol{x}_{mj} \in R_m.$$

Função de verossimilhança

Seja T_{mj} o tempo de falha para o j-ésimo indivíduo na região R_m e C_{mj} o tempo da censura. O tempo observado, é dado por $Y_{mj} = \min\{T_{mj}, C_{mj}\}$. A variável indicadora de censura δ_{mj} é definida sendo $\delta_{mj} = 1$ se $Y_{mj} = T_{mj}$, e $\delta_{mj} = 0$ caso contrário. Sendo que $h_0(t_{mj})$ é considerado como parâmetro *nuisance*, faz-se uso da verossimilhança parcial de Cox (Cox, 1972) dada por

$$L(\boldsymbol{\mathcal{T}}, \boldsymbol{\theta}|\mathcal{D}) = \prod_{m=1}^{M} \prod_{j=1}^{n_m} \left\{ \frac{\theta_m}{A_{mj}} \right\}^{\delta_{mj}}, \tag{5.2}$$

em que $A_{mj} = \sum_{l \in R(y_{mj})} \theta_m$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M)^\top$. $R_{(y_{mj})}$ é o conjunto de indivíduos em risco no tempo y_{mj}, i.e., $R_{(y_{mj})} = \{l : y_l \geq y_{mj}\}$. Para simplificar, considere-se uma covariável no modelo em que $M = 2$ no espaço preditor $\boldsymbol{\mathcal{X}}$, logo a função de verossimilhança dada em (5.2) pode ser reescrita na forma

$$L(\boldsymbol{\mathcal{T}}, \boldsymbol{\theta}|\mathcal{D}) = \prod_{j=1}^{n_1} \left\{ \frac{\theta_1}{A_{1j}} \right\}^{\delta_{1j}} \prod_{j=1}^{n_2} \left\{ \frac{\theta_2}{A_{2j}} \right\}^{\delta_{2j}}.$$

Pode-se observar que, no conjunto de risco de y_{mj}, existem indivíduos que pertencem à região R_1 ou R_2, em seguida, seja $n_{mj}^{R_m}$ que denota o número de indivíduos que pertencem a região R_m, $m = 1, 2$ e portanto A_{1j} é dada por

$$A_{1j} = \theta_1 n_{1j}^{R_1} + \theta_2 n_{1j}^{R_2}.$$

Considerando-se uma abordagem bayesiana, é necessário adotar uma distribuição *a priori* para os parâmetros locais θ_m. Neste caso, considera-se a distribuição gama para θ_m,

$$\theta_m \sim \text{Ga}(a_0, b_0),$$

em que a_0, b_0 são hiperparâmetros especificados. A distribuição *a posteriori* de $(\boldsymbol{\mathcal{T}}, \boldsymbol{\theta})$ é dada por

$$p(\boldsymbol{\mathcal{T}}, \boldsymbol{\theta}|\mathcal{D}) \propto \prod_{j=1}^{n_1} \left\{ \frac{\theta_1}{A_{1j}} \right\}^{\delta_{1j}} \prod_{j=1}^{n_2} \left\{ \frac{\theta_2}{A_{2j}} \right\}^{\delta_{2j}} p(\boldsymbol{\mathcal{T}}, \theta).$$

A distribuição *a posteriori* não é analiticamente tratável e, em seguida, usa-se o método MCMC para gerar amostras da distribuição *a posteriori*. Observa-se que $p(\boldsymbol{\mathcal{T}}, \boldsymbol{\theta}|\mathcal{D})$ pode ser definida sendo

$$p(\boldsymbol{\mathcal{T}}, \boldsymbol{\theta}|\mathcal{D}) = p(\boldsymbol{\theta}|\boldsymbol{\mathcal{T}}, \mathcal{D}) p(\boldsymbol{\mathcal{T}}|\mathcal{D}),$$

em que $p(\boldsymbol{\mathcal{T}}|\mathcal{D})$ é definida como $p(\boldsymbol{\mathcal{T}}|\mathcal{D}) \propto L(\boldsymbol{\mathcal{T}}|\mathcal{D}) p(\boldsymbol{\mathcal{T}})$, em que a verossimilhança marginal para $\boldsymbol{\mathcal{T}}$ é dada por

$$\begin{aligned} L(\boldsymbol{\mathcal{T}}|\mathcal{D}) &= \int L(\boldsymbol{\mathcal{T}}, \boldsymbol{\theta}|\mathcal{D}) p(\boldsymbol{\theta}|\boldsymbol{\mathcal{T}}) \mathrm{d}\boldsymbol{\theta} \\ &= \int \prod_{j=1}^{n_1} \left\{ \frac{\theta_1}{\theta_1 n_{1j}^{R_1} + \theta_2 n_{1j}^{R_2}} \right\}^{\delta_{1j}} p(\theta_1) \mathrm{d}\theta_m \int \prod_{j=1}^{n_2} \left\{ \frac{\theta_2}{\theta_1 n_{2j}^{R_1} + \theta_2 n_{2j}^{R_2}} \right\}^{\delta_{2j}} p(\theta_2) \mathrm{d}\theta_2. \quad (5.3) \end{aligned}$$

A integração dos parâmetros locais em (5.3) não pode ser feita em forma analítica e pode ser usada integração numérica para aproximar as integrais. Não obstante, tem-se parâmetros no denominador em (5.3) e isto leva a que a integração numérica não seja eficiente. Para contornar esse problema foram introduzidos variáveis latentes $\boldsymbol{z} = (z_1, \ldots, z_m)^\top$ na distribuição *a posteriori* $p(\boldsymbol{\mathcal{T}}, \boldsymbol{\theta}|\mathcal{D})$, assim a distribuição *a posteriori* conjunta para $(\boldsymbol{\mathcal{T}}, \boldsymbol{\theta}, \boldsymbol{z})$ é dada por

$$p(\boldsymbol{\mathcal{T}}, \boldsymbol{\theta}, \boldsymbol{z}|\mathcal{D}) \propto \exp\left(\theta_1 \sum_{j=1}^{n_1} \delta_{1j} - \sum_{j=1}^{n_1} z_{1j} A_{1j}^{\delta_{1j}} \right) \exp\left(\theta_2 \sum_{j=1}^{n_2} \delta_{2j} - \sum_{j=1}^{n_2} z_{2j} A_{2j}^{\delta_{2j}} \right) p(\boldsymbol{\mathcal{T}}, \boldsymbol{\theta}).$$

A condicional completa para z_{mj} (variável latente) é dada por

$$z_{mj}|\boldsymbol{\mathcal{T}}, \boldsymbol{\theta} \sim \text{Exp}(A_{mj}^{\delta_{mj}}), \quad m = 1, 2 \quad j = 1, \ldots, n_m.$$

Para obter as outras condicionais completas foi usada a relação

$$p(\boldsymbol{\mathcal{T}}, \boldsymbol{\theta}|\boldsymbol{z}, \mathcal{D}) = p(\boldsymbol{\theta}|\boldsymbol{\mathcal{T}}, \boldsymbol{z}, \mathcal{D}) p(\boldsymbol{\mathcal{T}}|\boldsymbol{z}, \mathcal{D}).$$

Assim, as condicionais completas para θ_1 e θ_2 são dadas por

$$\theta_1|\boldsymbol{\mathcal{T}}, \boldsymbol{z}, \mathcal{D} \sim \text{Ga}\left\{a_0, b_0 + \sum_{j=1}^{n_1}\left(z_{1j}\delta_{1j}n_{1j}^{R_1} - \delta_{1j}\right) + \sum_{j=1}^{n_2} z_{2j}\delta_{2j}n_{2j}^{R_1}\right\}$$
$$\theta_2|\boldsymbol{\mathcal{T}}, \boldsymbol{z}, \mathcal{D} \sim \text{Ga}\left\{a_0, b_0 + \sum_{j=1}^{n_1} z_{1j}\delta_{1j}n_{1j}^{R_2} + \sum_{j=1}^{n_2}\left(z_{2j}\delta_{2j}n_{2j}^{R_2} - \delta_{2j}\right)\right\}.$$

A distribuição condicional completa de $p(\boldsymbol{\mathcal{T}}|\boldsymbol{z}, \mathcal{D})$ é dada por

$$p(\boldsymbol{\mathcal{T}}|\boldsymbol{z}, \mathcal{D}) \propto c_1^* c_2^* c_3^*, \tag{5.4}$$

em que c_1^*, c_1^*, e c_1^* são dadas por

$$c_1^* = b_0^{a_0}\left\{\sum_{j=1}^{n_1}\left(z_{1j}\delta_{1j}n_{1j}^{R_1} - \delta_{1j}\right) + \sum_{j=1}^{n_2} z_{2j}\delta_{2j}n_{2j}^{R_1} + b_0\right\}^{-a_0},$$
$$c_2^* = b_0^{a_0}\left\{\sum_{j=1}^{n_1} z_{1j}\delta_{1j}n_{1j}^{R_2} + \sum_{j=1}^{n_2}\left(z_{2j}\delta_{2j}n_{2j}^{R_2} - \delta_{2j}\right) + b_0\right\}^{-a_0}$$

e

$$c_3^* = \exp\left\{-\sum_{j=1}^{n_1}(1-\delta_{1j})z_{1j} - \sum_{j=1}^{n_2}(1-\delta_{2j})z_{2j}\right\}.$$

Nota-se que, $p(\boldsymbol{\mathcal{T}}|\boldsymbol{z}, \mathcal{D})$ não depende dos parâmetros locais, $\boldsymbol{\theta}$.

5.2.2 Distribuição Gompertz defeituosa

Uma distribuição é chamada de defeituosa se sua função densidade é imprópria para alguns valores dos parâmetros. Neste sentido, exemplos de distribuições defeituosas são a distribuição Gompertz e gaussiana inversa.

A função densidade da distribuição Gompertz é dada por

$$f(t|\lambda, \alpha) = \lambda e^{\alpha t}\exp\left\{-(\lambda/\alpha)\left(e^{\alpha t} - 1\right)\right\}, \quad t \geq 0$$

em que $\lambda > 0$ e $\alpha > 0$. A função de sobrevivência para o modelo Gompertz é dada por

$$S(t|\lambda, \alpha) = \exp\left\{-(\lambda/\alpha)\left(e^{\alpha t} - 1\right)\right\}. \tag{5.5}$$

Nota-se em (5.5) que se $\alpha < 0$ tem-se que a distribuição Gompertz é imprópria assim

$$\lim_{t\to\infty} S(t|\lambda, \alpha) = e^{\lambda/\alpha}.$$

Sendo que a distribuição Gompertz é uma distribuição defeituosa, foi considerada para a modelagem de dados de sobrevivência com fração de cura como pode ser visto em Cantor & Shuster (1992) e Gieser *et al.* (1998).

Uma proposta futura de trabalho é considerar uma extensão do modelo Gompertz defeituoso baseado no modelo de partição bayesiana. Nessa extensão, pode-se considerar como parâmetro local o parâmetro α em cada região da tesselação, porém o amostrador MCMC tem que ser modificado.

Apêndice A

Gráficos da simulação MCMC do modelo MPB para o conjunto de dados de melanoma.

Modelo MPBLg

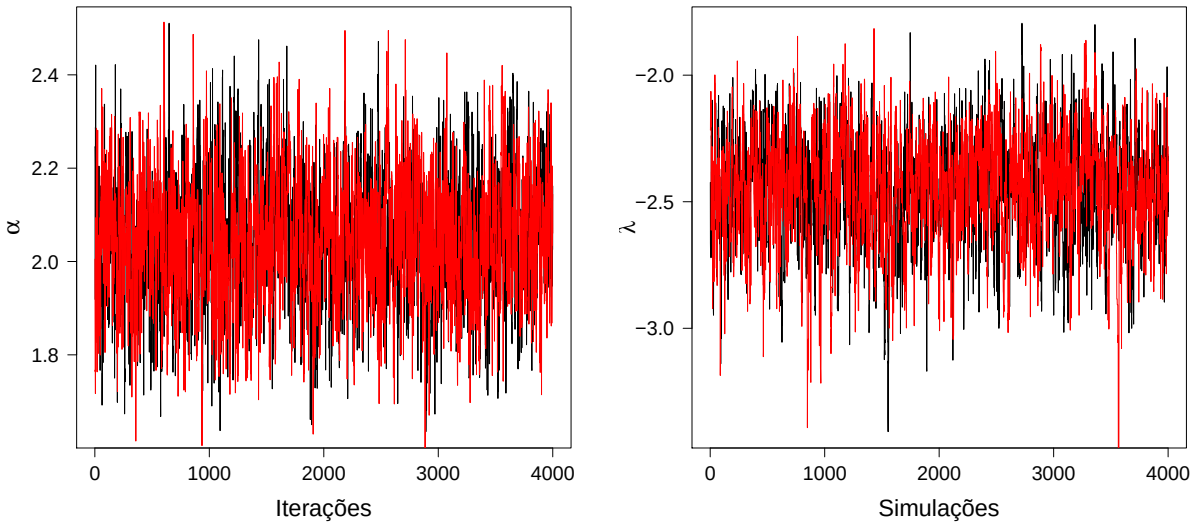


Figura A.1: Histórico da seqüência de iterações dos parâmetros da distribuição Weibull do modelo MPBLg.

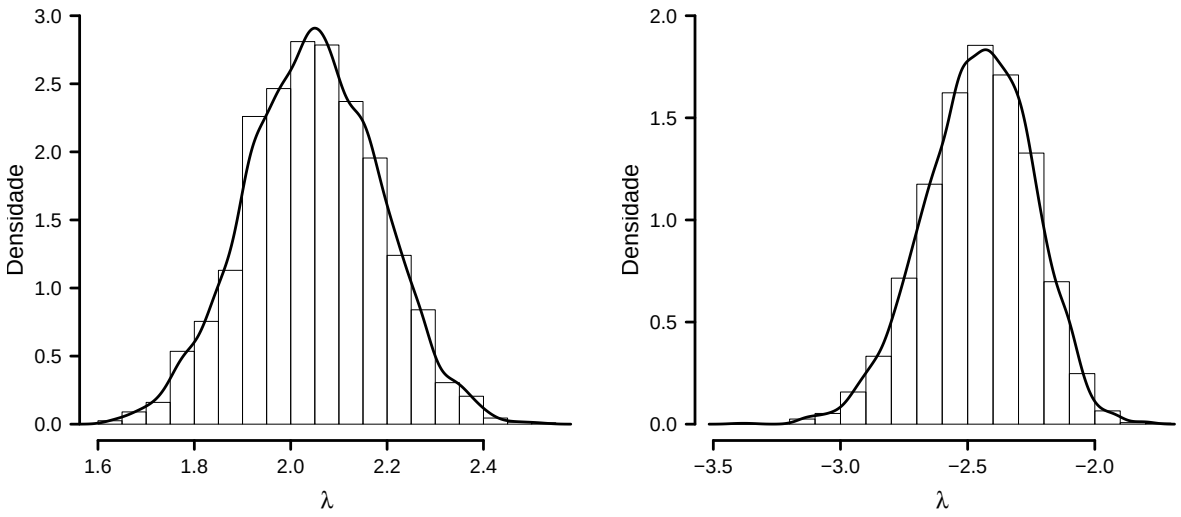


Figura A.2: Densidades marginais *a posteriori* aproximadas para os parâmetros da distribuição Weibull do modelo MPBLg.

Modelo MPBGeo

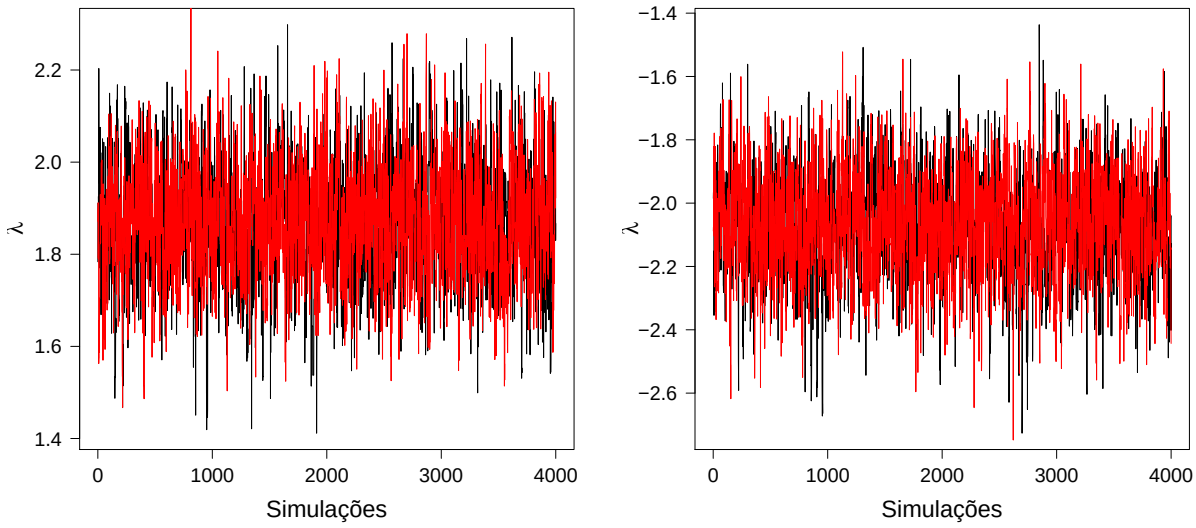


Figura A.3: Histórico da seqüência de iterações dos parâmetros da distribuição Weibull do modelo MPBGeo.

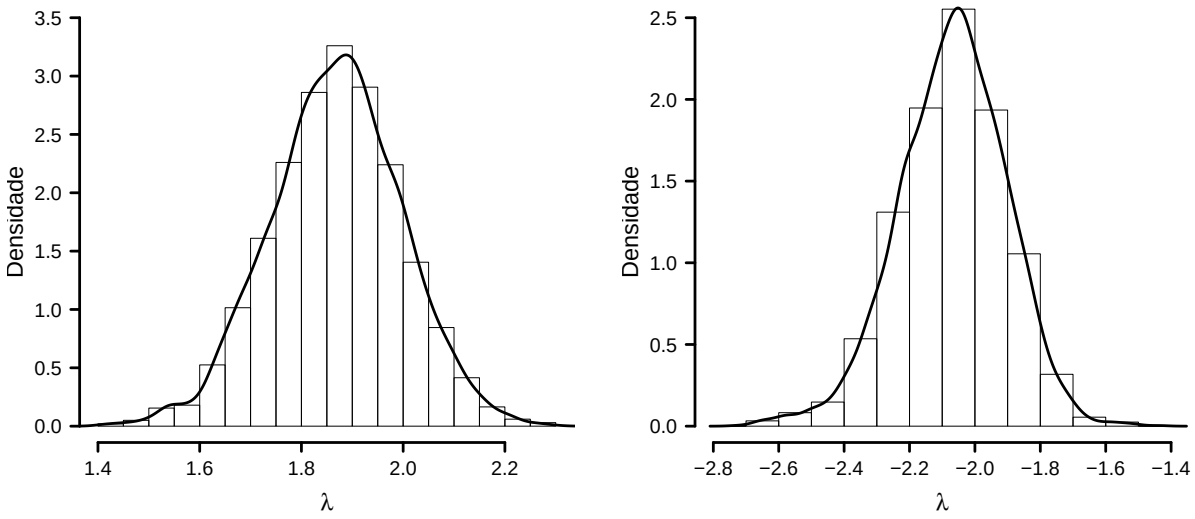


Figura A.4: Densidades marginais *a posteriori* aproximadas para os parâmetros da distribuição Weibull do modelo MPBGeo.

Modelo MPBPoi

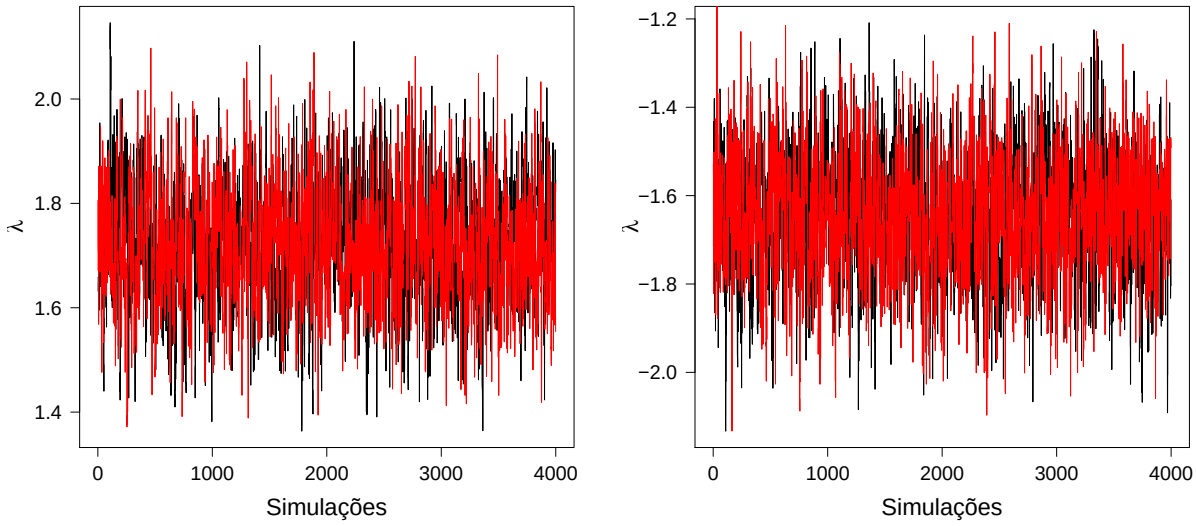


Figura A.5: Histórico da seqüência de iterações dos parâmetros da distribuição Weibull do modelo MPBPoi.

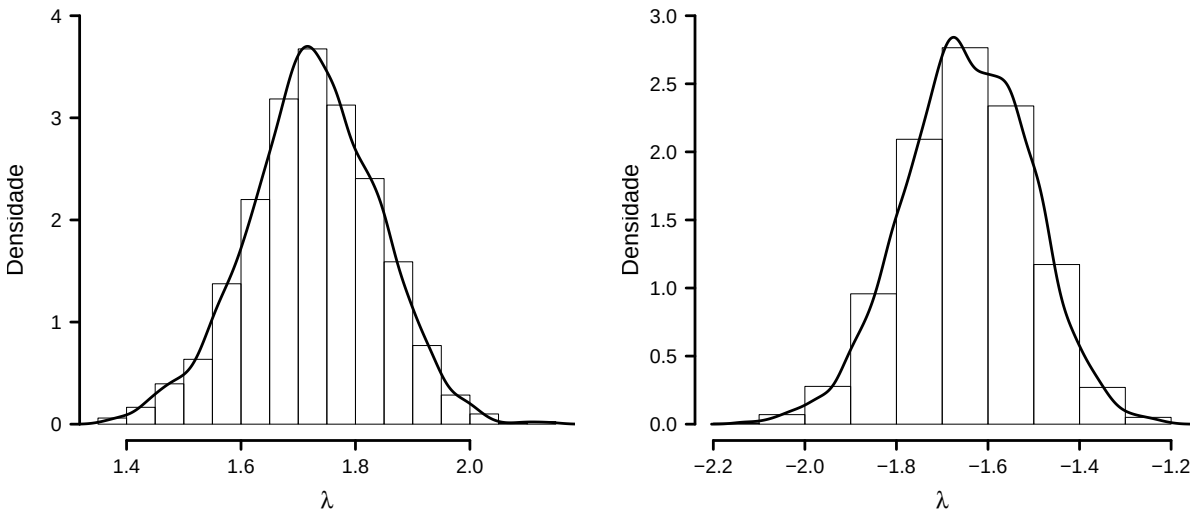


Figura A.6: Densidades marginais *a posteriori* aproximadas para os parâmetros da distribuição Weibull do modelo MPBPoi.

Modelo MPBBi com $K = 10$

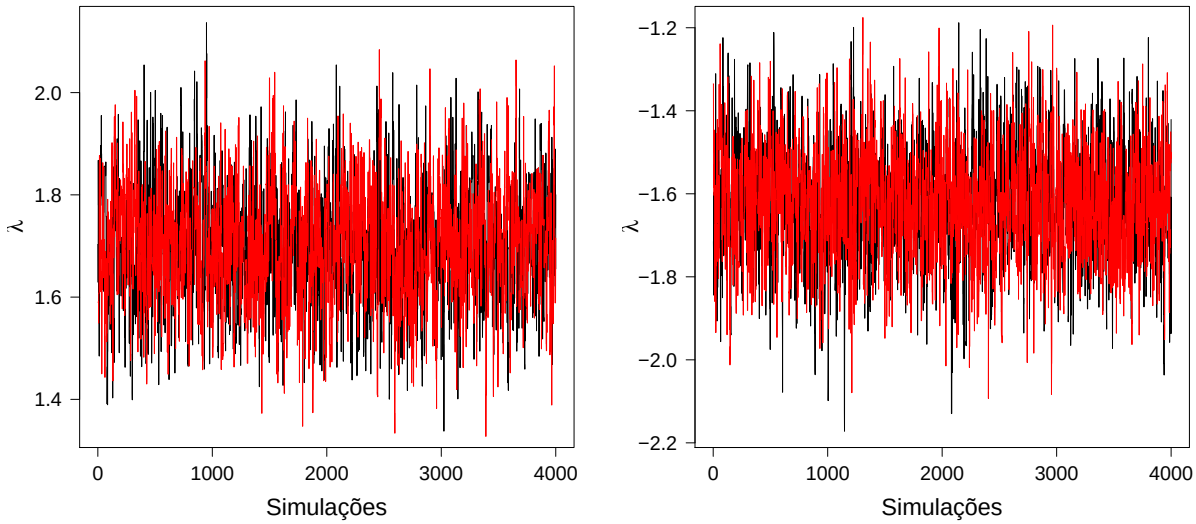


Figura A.7: Histórico da seqüência de iterações dos parâmetros da distribuição Weibull do modelo MPBBi com $K = 10$.

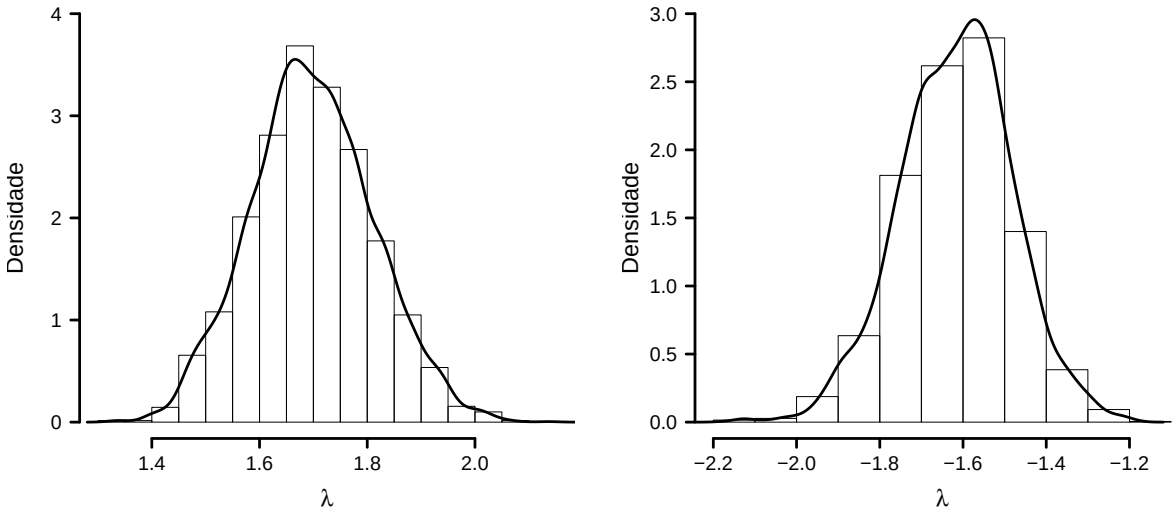


Figura A.8: Densidades marginais *a posteriori* aproximadas para os parâmetros da distribuição Weibull do modelo MPBBi com $K = 10$.

Apêndice B

Gráficos da simulação MCMC do modelo MPB para o conjunto de dados de leucemia.

Modelo MPBLg

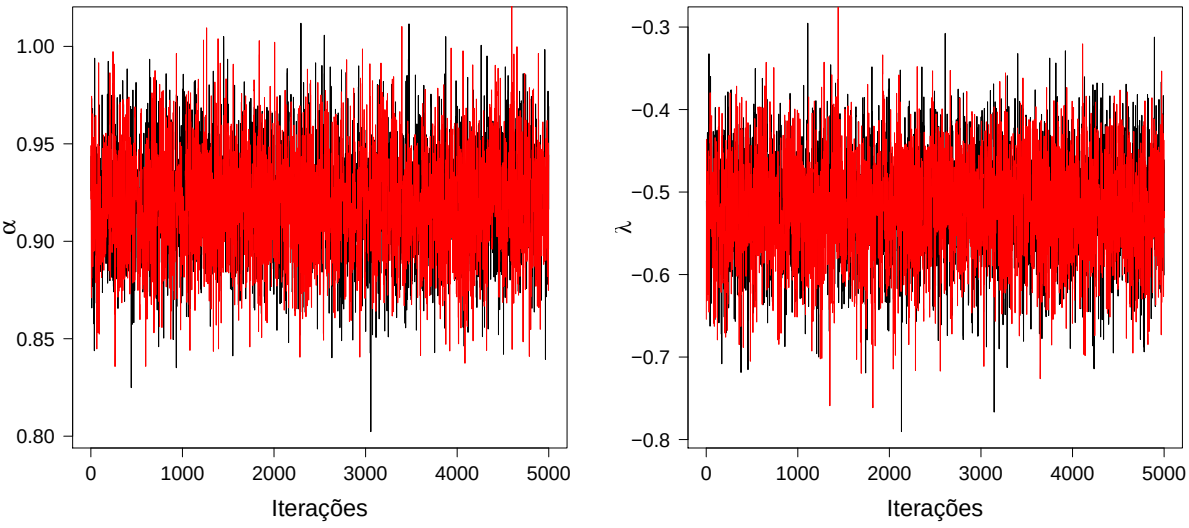


Figura B.1: Histórico da seqüência de iterações dos parâmetros da distribuição Weibull do modelo MPBLg.

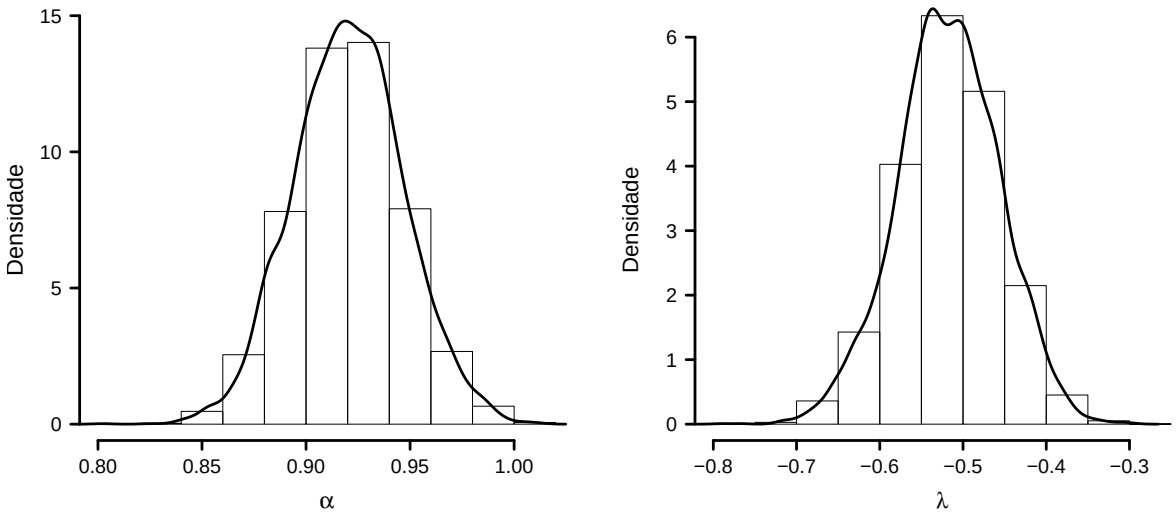


Figura B.2: Densidades marginais *a posteriori* aproximadas para os parâmetros da distribuição Weibull do modelo MPBLg.

Modelo MPBGeo

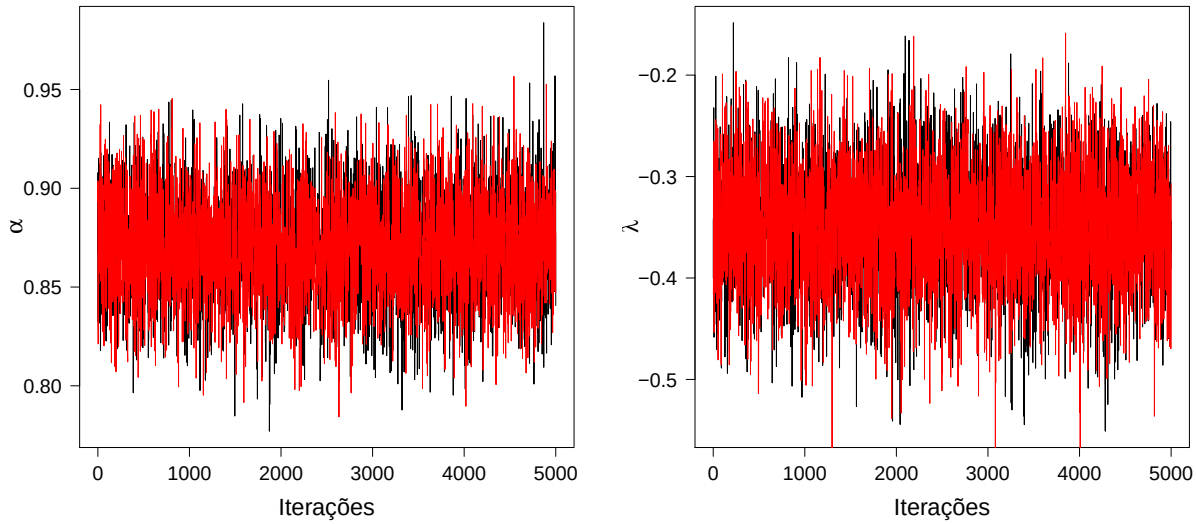


Figura B.3: Histórico da seqüência de iterações dos parâmetros da distribuição Weibull do modelo MPBGeo.

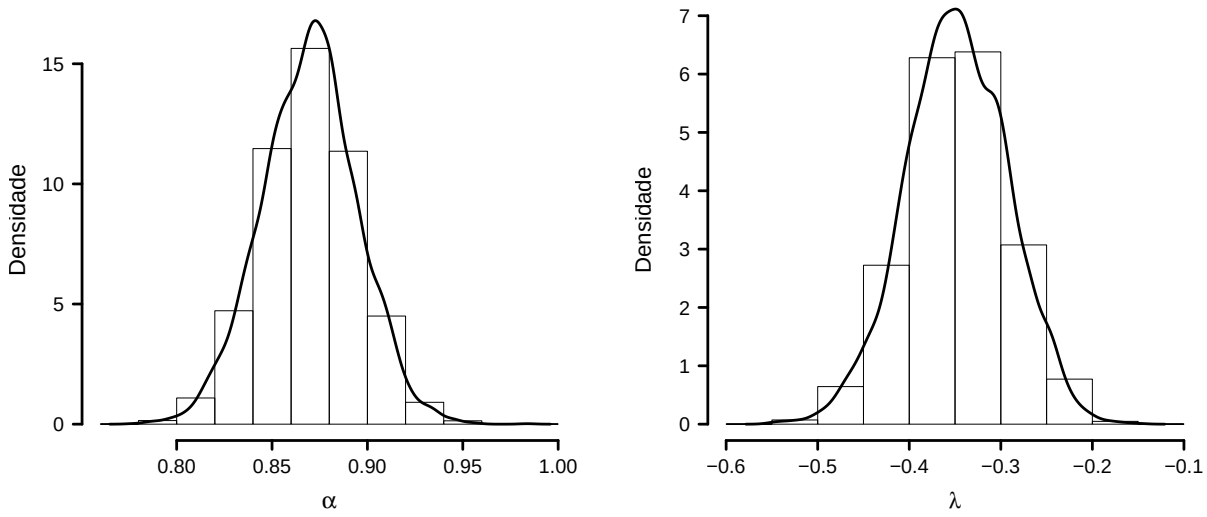


Figura B.4: Densidades marginais *a posteriori* aproximadas para os parâmetros da distribuição Weibull do modelo MPBGeo.

Modelo MPBPoi

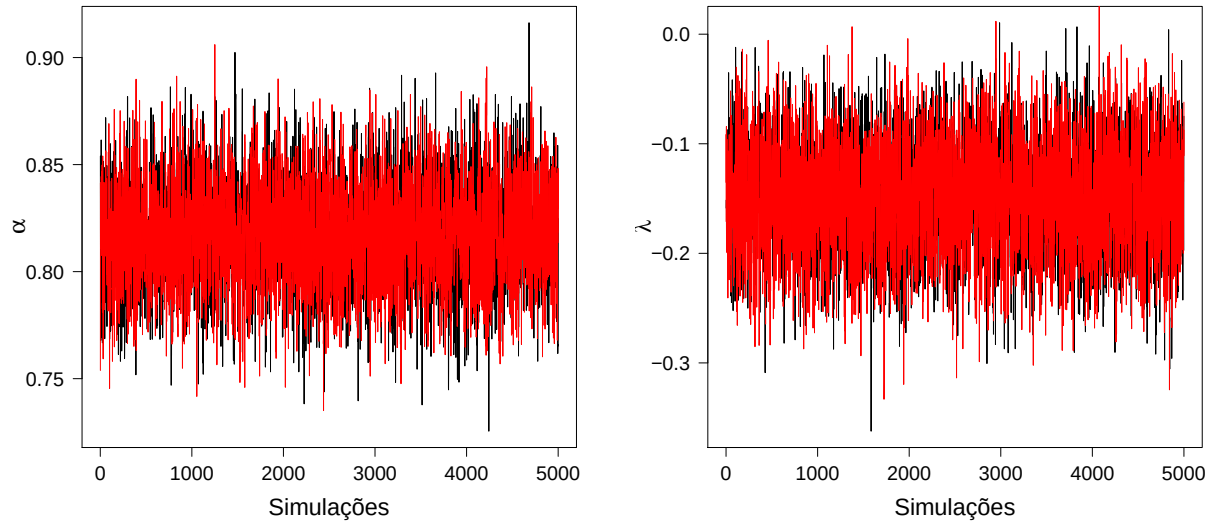


Figura B.5: Histórico da seqüência de iterações dos parâmetros da distribuição Weibull do modelo MPBPoi.

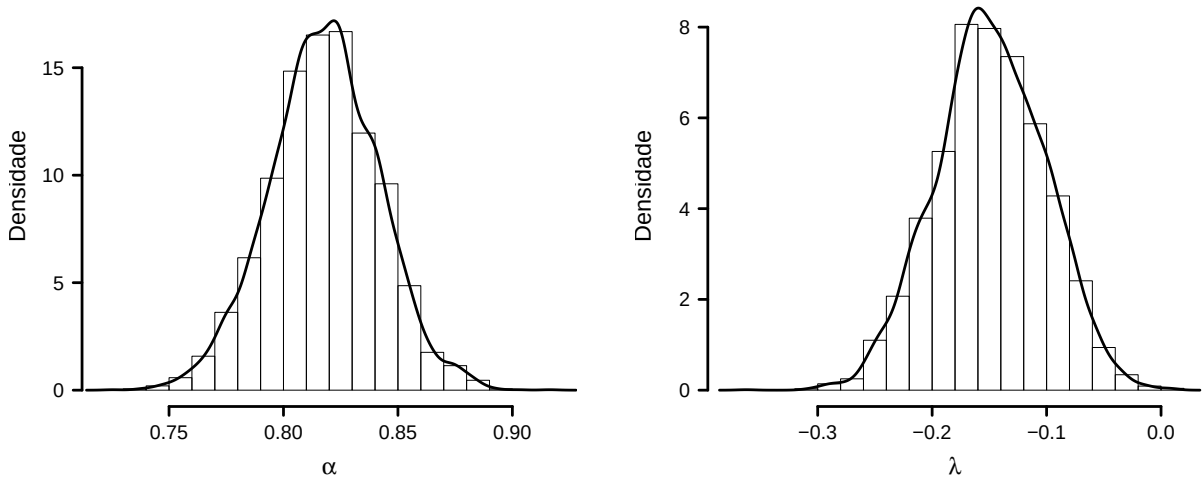


Figura B.6: Densidades marginais *a posteriori* aproximadas para os parâmetros da distribuição Weibull do modelo MPBPoi.

Modelo MPBBi com $K = 30$

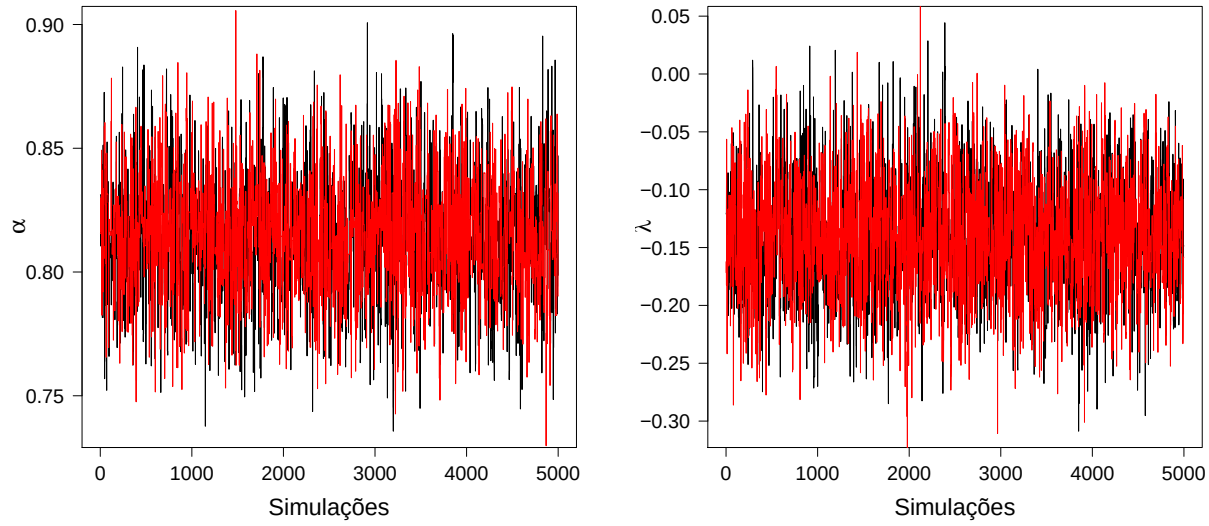


Figura B.7: Histórico da seqüência de iterações dos parâmetros da distribuição Weibull do modelo MPBBi com $K = 30$.

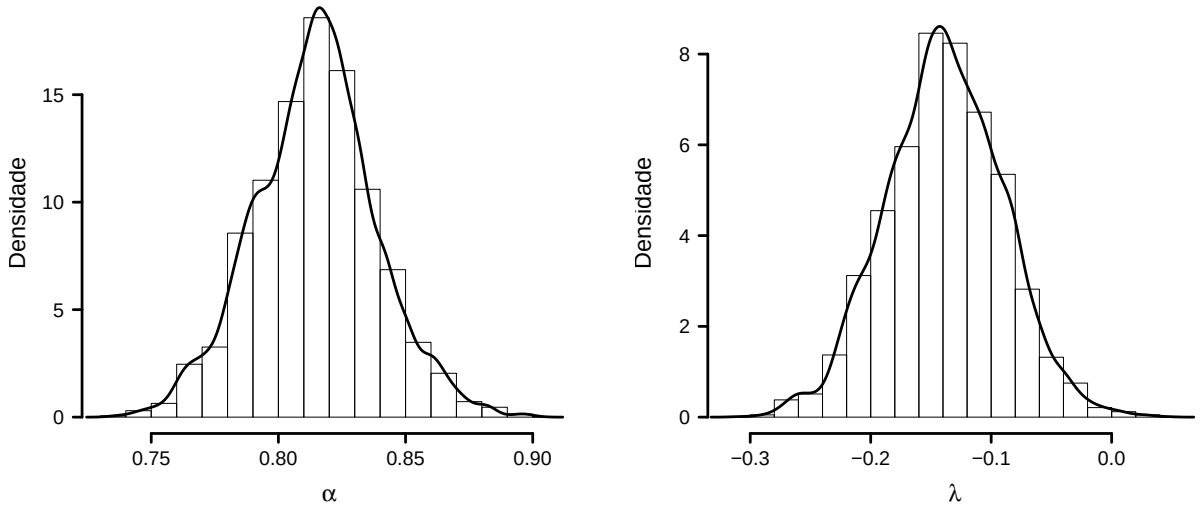


Figura B.8: Densidades marginais *a posteriori* aproximadas para os parâmetros da distribuição Weibull do modelo MPBBi com $K = 30$.

Referências

Aalen, O. O. (1978). Statistical inference for a family of counting processes. *The Annals of Statistics*, **6**, 701–726. 10

Barry, D. & Hartigan, J. A. (1992). Product partition models for change point problems. *The Annals of Statistics*, **20**, 260–279. 30

Barry, D. & Hartigan, J. A. (1993). A Bayesian analysis for change point problems. *Journal of the American Statistical Association*, **80**, 309–319. 5

Berkson, J. & Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, **47**, 501–515. 3, 4

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, **11**, 15–53. 2, 3, 4

Brooks, S., Gelman, A., Jones, G. L. & Meng, X.-L., editors (2011). *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC, Boca Raton, FL. 46

Cancho, V. G., Rodrigues, J. & de Castro, M. (2011). A flexible model for survival data with a cure rate: a Bayesian approach . *Journal of Applied Statistics*, **38**, 57 – 70. 4

Cantor, A. B. & Shuster, J. J. (1992). Parametric versus non-parametric methods for estimating cure rates based on censored survival data. *Statistics in Medicine*, **11**, 931–937. 83

Chen, B. E., Jiang, W. & Tu, D. (2014). A hierarchical Bayes model for biomarker subset effects in clinical trials. *Computational Statistics and Data Analysis*, **71**, 324–334. 80

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. 2, 4, 6, 15, 17

Chen, M.-H., Shao, Q. M. & Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer, New York. 47, 52

Chipman, H. A., George, E. I. & McCulloch, R. E. (1998). Bayesian CART model search. *Journal of the American Statistical Association*, **93**, 935–948. 31, 34, 41

Contal, C. & O'Quigley, J. (1999). An application of changepoint methods in studying the efect of age on survival in breast cancer. *Computational Statistics and Data Analysis*, **30**, 253–270. 80

Cooner, F., Banerjee, S., Carlin, B. P. & Sinha, D. (2007). Flexible cure rate modeling under latent activation schemes. *Journal of the American Statistical Association*, **102**, 560–572. 15, 18

Cox, D. R. (1972). Regression models and life-tables (with discussion) . *Journal of the Royal Statistical Society. Series B (Methodological)*, **34**, 187–220. 1, 80

de Castro, M., Cancho, V. G. & Rodrigues, J. (2009). A Bayesian long-term survival model parametrized in the cured fraction. *Biometrical Journal*, **51**, 443–55. 4

Denison, D. G. T. & Holmes, C. C. (2001). Bayesian partitioning for estimating disease risk. *Biometrics*, **57**, 143–149. 5, 32

Denison, D. G. T., Mallick, B. K. & Smith, A. F. M. (1998). A Bayesian CART algorithm. *Biometrika*, **85**, 363–377. 31, 41

Denison, D. G. T., Adams, N. M., Holmes, C. C. & Hand, D. J. (2002a). Bayesian partition modelling. *Computational Statistics and Data Analysis*, **38**, 475–485. 78

Denison, D. G. T., Holmes, C. C., Mallick, B. K. & Smith, A. F. M. (2002b). *Bayesian Methods for Nonlinear Classification and Regression*. Wiley, Chichester. 31, 35, 39

Farewell, V. T. (1977). A model for binary variable with time-censored observations. *Biometrika*, **38**, 43–46. 3

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, **38**, 1041–1046. 3

Farewell, V. T. & Sprott, D. (1986). Mixture models in survival analysis: are they worth the risk? *The Canadian Journal of Statistics*, **14**, 257–262. 2, 3

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, **19**, 1–67. 29, 30

Gail, M. H., Santner, T. J. & Brown, C. C. (1980). An analysis of comparative carcinogenesis experiments based on multiple times to tumor. *Biometrics*, **36**, 2. 15

Gelman, A. & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–472. 53, 66

Gieser, P. W., Chang, M. N., Rao, P. V., Shuster, J. J. & Pullen, J. (1998). Modelling cure rates using the Gompertz model with covariate information. *Statistics in Medicine*, **17**, 831–839. 83

Gilks, W. R. & Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **41**, 337–348. 45, 51

Giudici, P., Knorr-Held, L. & Rasser, G. (2000). Modelling categorical covariates in Bayesian disease mapping by partition structures . *Statistics in Medicine*, **19**, 2579–2593. 40

Goldman, A. I. (1984). Survivorship analysis when cure is a possibility: A Monte Carlo study. *Statistics in Medicine*, **3**, 153–163. 3

Gonzales, J. F. B., Tomazella, V. & Taconelli, J. P. (2013). Estimação paramétrica do modelo de mistura com fragilidade gama na presença de covariáveis. *Rev. Bras. Biom*, **31**, 233–247. 7, 29

Gonzales, J. F. B. G., Tomazella, V., de Castro, M. & Louzada, F. (2012). A Bayesian partition modelling approach for geometric cure rate survival models . Technical report, Relatório Técnico do DEs - Teoría & Métodos 251, São Carlos, Brasil. ISSN 0104-0499. 7

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732. 5, 37

Green, P. J. (2003). Trans-dimensional Markov chain Monte Carlo. In P. Green, N. Hjort, & S. Richardson, editors, *Highly Structured Stochastic Systems*, pages 179–198. Oxford University Press, Oxford. 39

Greenhouse, J. B. & Wolfe, R. A. (1984). A competing risks derivation of a mixture model for the analysis of survival data. *Communication in Statistics - Theory and Methods*, **13**, 3133–3154. 3

Gu, Y., Sinha, D. & Banerjee, S. (2011). Analysis of cure rate survival data under proportional odds model. *Lifetime Data Analysis*, **17**, 123–134. 4

Han, C. & Carlin, B. P. (2001). Markov Chain Monte Carlo Methods for Computing Bayes Factors: A Comparative Review. *Journal of the American Statistical Association*, **96**, 1122–1132. 37

Hanin, L. G. (2001). Iterated birth and death process as a model of radiation cell survival . *Mathematical Biosciences*, **169**, 89–107. 4

Hartigan, J. A. (1990). Partition models . *Communications in Statistics*, **19**, 2745–2756. 5, 30

Hegarty, A. & Barry, D. (2008). Bayesian disease mapping using product partition models. *Statistics in Medicine*, **27**, 3868–3893. 5

Heikkinen, J. (1998). Curve and surface estimation using dynamic step functions. In D. K. Dey, editor, *Practical Nonparametric and Semiparametric Bayesian Statistics, no. 133 in Lecture Notes in Statistics, chap. 14* , pages 255–272, New York. Springer-Verlag. 5

Heikkinen, J. & Arjas, E. (1998). Non-parametric Bayesian estimation of a spatial Poisson intensity. *Scandinavian Journal of Statistics*, **25**, 435–450. 31

Heikkinen, J. & Arjas, E. (1999). Modeling a Poisson forest in variable elevations: a nonparametric Bayesian approach. *Biometrics*, **55**, 738–745. 31

Hoggart, C. & Griffin, J. E. (2001). A Bayesian partition model for customer attrition. In E. I. George, editor, *Bayesian Methods with Applications to Science, Policy, and Official Statistics(Selected Papers from ISBA 2000)*, pages 61–70, Creta,Greece. International Society for Bayesian Analysis, Proceedings of the the Sixth World Meeting of the International Society for Bayesian Analysis. i, ii, 3, 37, 40, 44, 78

Holmes, C. C. & Mallick, B. K. (2003). Generalized nonlinear modeling with multivariate free-knot regression splines. *Journal of the American Statistical Association*, **98**, 352–368. 29

Holmes, C. C., Denison, D. G. T. & Mallick, B. K. (1999). Bayesian partitioning for classification and regression. Technical report, Department of Mathematics, Imperial College. 5, 7, 31, 32, 37, 78

Holmes, C. C., Denison, D. G. T., Ray, S. & Mallick, B. K. (2005). Bayesian prediction via partitioning. *Journal of Computational and Graphical Statistics*, **14**, 811–830. 5, 7, 31, 32, 34, 37, 39, 43, 78, 79

Hopcroft, P. O., Gallagher, K. & Pain, C. (2009). A Bayesian partition modelling approach to resolve spatial variability in climate records from borehole temperature inversion. *Geophysical Journal International*, **178**, 651–666. 78

Ibrahim, J. G., Chen, M.-H. & Sinha, D. (2001a). Bayesian semiparametric models for survival data with a cure fraction. *Biometrics*, **57**, 383–388. 4, 5

Ibrahim, J. G., Chen, M.-H. & Sinha, D. (2001b). *Bayesian Survival Analysis.*. Springer, New York. 10, 45, 51

Jensen, U. & Lütkebohmert, C. (2008). A Cox-type regression model with change-points in the covariates. *Lifetime Data Anal*, **14**, 267–285. 80

Johnson, N. L., Kemp, A. W. & Kotz, S. (2005). *Univariate Discrete Distributions* . John Wiley & Sons, Hoboken, NJ, third edition. 13

Kalbfleisch, J. D. & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, Hoboken, NJ, third edition. 2

Kaplan, E. L. & Meier, P. (1958). Nonparametric estimation from incomplete observations . *Journal of the American Statistical Association*, **53**, 457–481. 1

Kemp, A. W. (1981). Efficient generation of logarithmically distributed pseudo-random variables . *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **30**, 249–253. 51

Kim, H. M., Mallick, B. K. & Holmes, C. C. (2005). Analyzing nonstationary spatial data using piecewise Gaussian processes. *Journal of the American Statistical Association*, **100**, 653–668. 36, 37

Kim, S., Chen, M.-H., Dey, D. K. & Gamerman, D. (2007). Bayesian dynamic models for survival data with a cure fraction. *Lifetime Data Analysis*, **13**, 17–35. 5

Kirkwood, J. M., Ibrahim, J. G., Sondak, V. K., Richards, J., Flaherty, L. E., Ernstoff, M. S., Smith, T. J., Rao, U., Steele, M. & Blum, R. H. (2000). High- and low-dose interferon alfa-2b in high-risk melanoma: First analysis of intergroup trial e1690/s9111/c9190. *Journal of Clinical Oncology*, **18**, 2444–2458. 10

Kosambi, D. D. (1949). Characteristic properties of series distributions. *Proceedings of the National Institute for Science, India*, **15**, 109–113. 13

Kuk, A. Y. C. & Chen, C. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, **79**, 531–541. 3, 6

Lambert, P., Thompson, J., Weston, C. & Dickman, P. (2007). Estimating and modeling the cure fraction in population-based cancer survival analysis. *Biostatistics*, **8**, 576–594. 2

Lawless, J. (2002). *Statiscal Models and Methods for Lifetime Data*. Wiley, New York,NY, second edition. 1

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. 47

Louzada, F., de Castro, M., Tomazella, V. & Gonzales, J. F. B. (2014). Modeling categorical covariates for lifetime data in the presence of cure fraction by Bayesian partition structures. *Journal of Applied Statistics*, **41**, 622–634. 7

Maller, R. A. & Zhou, S. (1996). *Survival Analysis with Long-Term Survivors.*. Wiley, New York,NY. 3

McCullagh, P. & Yang, J. (2008). How many clusters? *Bayesian Analysis*, **3**, 101–120. 5, 38

Meeker, W. Q. & Escobar, L. A. (1998). *Statistical Methods for Reliability Data.*. Wiley, New York,NY. 3

Moolgavkar, S. H., Luebeck, E. G. & de Gunst, M. (1990). Two-Mutation Model for Carcinogenesis: Relative Roles of Somatic Mutations and Cell Proliferation in Determining Risk. In *Scientific Issues in Quantitative Cancer Risk Assessment*, pages 136–152. Boston: Birkhauser. 18

Muller, P. & Quintana, F. A. (2010). Random partition models with regression on covariates. *Journal of Statistical Planning and Inference*, **140**, 2801–2808. 5

Nelson, W. B. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, **14**, 945–966. 10

Noack, A. (1950). A class of random variables with discrete distributions. *The Annals of Mathematical Statistics*, **21**, 127–132. 13

Okabe, A., Boots, B., Sugihara, K. & Chiu, S. N. (2000). *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. Wiley, Chichester. 32

Peng, Y. & Dear, K. B. G. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics*, **56**, 237–243. 2, 3, 6

Peng, Y., Dear, K. B. G. & Denham, J. W. (1998). A generalized F mixture model for cure rate estimation. *Statistics in Medicine*, **17**, 813–830. 3

Putter, H. (2011). *dynpred: Companion package to "Dynamic Prediction in Clinical Survival Analysis"*. R package version 0.1.1. 8

Quintana, F. A. & Iglesias, P. L. (2003). Bayesian clustering and product partition models. *Journal of the Royal Statistical Society Series B*, **65**, 557–574. 5

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 21, 79

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. 4, 7, 12, 20, 78

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. 4

Rodrigues, J., de Castro, M., Balakrishnan, N. & Cancho, V. G. (2011). Destructive weighted Poisson cure rate models. *Lifetime Data Analysis*, **17**, 333–346. 4

Segal, M. R. (1988). Regression trees for censored data. *Biometrics*, **44**, 35–47. 5

Stephens, D. A. (1994). Bayesian retrospective multiple-changepoint identification. *Applied Statistics*, **43**, 159–178. 5

Stoyan, P. D., Kendall, D. W. S. & Mecke, J. (1995). *Stochastic Geometry and Its Applications*. John Wiley & Sons, Chichester, NY. 32

Sy, J. P. & Taylor, J. M. G. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics*, **56**, 227–236. 3

Tomazella, V. L. D., de Castro, M., Louzada-Neto, F. & Gonzales, J. F. B. (2012). Bayesian partition for Poisson cure rate survival models . Technical report, Relatório Técnico do DEs - Teoría & Métodos 252, São Carlos, Brasil. ISSN 0104-0499. 7

Tomazella, V. L. D., de Castro, M. & Gonzales, J. F. B. (2013). A flexible Bayesian partition modelling for long-term survival data . Technical report, Relatório Técnico do DEs - Teoría & Métodos 257, São Carlos, Brasil. ISSN 0104-0499. 7

Tong, E. N. C., Mues, C. & Thomas, L. C. (2012). Mixture cure models in credit scoring: If and when borrowers default. *European Journal of Operational Research*, **218**, 132–139. 4

Tsodikov, A. D., Ibrahim, J. G. & Yakovlev, A. Y. (2003). Estimating cure rates from survival data: An alternative to two-component mixture models. *Journal of the American Statistical Association*, **98**, 1063–1078. 4, 7, 12, 78

Yakovlev, A. Y. & Tsodikov, A. D. (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. World Scientific, Singapore. 4, 17

Yin, G. & Ibrahim, J. G. (2005). Cure rate models: A unified approach. *The Canadian Journal of Statistics*, **33**, 559–570. 5

Zhang, H. & Singer, B. H. (2010). *Recursive Partitioning and Applications*. Springer, New York, second edition. 5