

UNIVERSIDADE ESTADUAL PAULISTA

Programa de Pós-Graduação em Matemática Aplicada e Computacional

**Uma Classe de Modelos Bayesianos
Semiparamétricos de Transformação para
Dados de Sobrevivência com Fração de Cura**

Discente: Maria Luisa Scherer Belini

Orientador: Prof. Dr. Ricardo Puziol de Oliveira

UNIVERSIDADE ESTADUAL PAULISTA

Faculdade de Ciências e Tecnologia de Presidente Prudente

Programa de Pós-Graduação em Matemática Aplicada e Computacional

Uma Classe de Modelos Bayesianos Semiparamétricos de Transformação para Dados de Sobrevivência com Fração de Cura

Maria Luisa Scherer Belini

Orientadora: Prof. Dr. Ricardo Puziol de Oliveira

Dissertação apresentada ao Programa de Pós-Graduação em Matemática Aplicada e Computacional da Faculdade de Ciências e Tecnologia da UNESP para obtenção do título de Mestre em Matemática Aplicada e Computacional.

B431c	Belini, Maria Luisa Scherer Uma Classe de Modelos Bayesianos Semiparamétricos de Transformação para Dados de Sobrevida com Fração de Cura / Maria Luisa Scherer Belini. -- Presidente Prudente, 2026 64 p. Dissertação (mestrado) - Universidade Estadual Paulista (UNESP), Faculdade de Ciências e Tecnologia, Presidente Prudente Orientador: Ricardo Puziol de Oliveira 1. Análise de sobrevivência. 2. Fração de cura. 3. Inferência Bayesiana. 4. Modelos de transformação. 5. Modelos semiparamétrico. I. Título.
-------	---

CERTIFICADO DE APROVAÇÃO

TÍTULO DA DISSERTAÇÃO: Uma Classe de Modelos Bayesianos Semiparamétricos de Transformação para Dados de Sobrevida com Fração de Cura

AUTORA: MARIA LUISA SCHERER BELINI

ORIENTADOR: RICARDO PUZIOL DE OLIVEIRA

Aprovada como parte das exigências para obtenção do Título de Mestra em Matemática Aplicada e Computacional, pela Comissão Examinadora:

Documento assinado digitalmente
 **RICARDO PUZIOL DE OLIVEIRA**
Data: 24/09/2026 11:45:28-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. RICARDO PUZIOL DE OLIVEIRA (Participação Virtual)
Departamento de Estatística / UNESP / Câmpus de Presidente Prudente - FCT

Documento assinado digitalmente
 **FERNANDO ANTONIO MOALA**
Data: 24/09/2026 13:19:26-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. FERNANDO ANTONIO MOALA (Participação Virtual)
Departamento de Estatística / UNESP / Câmpus de Presidente Prudente - FCT

Documento assinado digitalmente
 **JORGE ALBERTO ACHCAR**
Data: 24/09/2026 09:25:56-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. JORGE ALBERTO ACHCAR (Participação Virtual)
FMRP/Departamento de Medicina Social / Universidade de São Paulo

Presidente Prudente, 01 de setembro de 2026.

*Aqueles que me ensinaram a acreditar no conhecimento como caminho de
transformação.*

Agradecimentos

A Deus, pela força e pela oportunidade de chegar até aqui.

Ao meu orientador, Prof. Dr. Ricardo Puziol de Oliveira, pela orientação, paciência, confiança e pelos ensinamentos que foram fundamentais para o desenvolvimento deste trabalho.

Ao Prof. Dr. Fernando Antonio Moala e ao Prof. Dr. Jorge Alberto Achcar, pelas contribuições e sugestões ao longo do processo de pesquisa.

Ao Programa de Pós-Graduação em Matemática Aplicada e Computacional da UNESP (POSMAC), ao qual estou vinculada, pela formação e pela estrutura oferecida ao longo do mestrado.

À Universidade Virtual do Estado de São Paulo (UNIVESP), pelo apoio financeiro concedido durante minha permanência no programa.

À minha família e a todos que, de alguma forma, contribuíram para a realização desta dissertação, deixo minha sincera gratidão.

*“ A verdadeira viagem plena de descobertas
não consiste em ver novas paisagens, mas
em ter um olhar sempre renovado.”*
Marcel Proust

Resumo

Modelos de sobrevivência com fração de cura são adequados a contextos em que uma parcela da população pode permanecer livre do evento de interesse em longo prazo, comportamento associado à presença de um platô na curva de sobrevivência. Esta dissertação propõe uma classe de modelos Bayesianos semiparamétricos de transformação do risco acumulado com fração de cura, na qual o componente de incidência representa a probabilidade de cura por meio de uma ligação *logit*, enquanto o componente de latência descreve o tempo até o evento entre os indivíduos suscetíveis. Foram consideradas transformações previamente especificadas, incluindo as estruturas de riscos proporcionais e chances proporcionais, além de uma transformação monótona adaptativa estimada a partir dos dados por meio de um gerador construído sobre funções de base *B-spline* cúbicas. O risco acumulado *baseline* foi representado de forma não paramétrica por incrementos positivos definidos nos tempos de evento observados, com regularização baseada em penalizações de passeio aleatório de segunda ordem. A inferência a posteriori foi conduzida por meio de um algoritmo Metropolis-within-Gibbs, combinando propostas de passeio aleatório multivariadas, atualizações MALA preconditionadas e etapas de Gibbs. O desempenho do modelo adaptativo foi avaliado em um estudo de simulação de Monte Carlo com tamanhos amostrais de 30, 50, 100 e 500 e níveis de censura de 10% e 30%. Os resultados mostraram baixos valores de viés para os coeficientes de latência, redução do RMSE com o aumento do tamanho amostral e recuperação da função de transformação e do risco acumulado *baseline*, enquanto os componentes de incidência e transformação apresentaram maior incerteza em alguns cenários. A aplicação foi realizada em uma coorte retrospectiva de pacientes com câncer de mama do Netherlands Cancer Institute. O modelo adaptativo apresentou o menor DIC entre os modelos Bayesianos avaliados e WAIC próximo ao menor valor obtido pelo modelo de riscos proporcionais. Entre as pacientes suscetíveis, o grau tumoral III foi associado à ocorrência mais precoce de eventos. A análise de sensibilidade às priors indicou estabilidade dos efeitos de regressão e das funções estimadas, enquanto a análise de correlação a posteriori evidenciou dependência entre alguns efeitos de incidência e latência. Os resultados mostram que a proposta permite estimar a transformação do risco a partir dos dados, mantendo a separação entre os mecanismos associados à cura e à sobrevivência dos indivíduos suscetíveis.

Palavras-Chave: *Análise de sobrevivência, fração de cura, inferência Bayesiana, modelos de transformação, modelos semiparamétricos.*

Abstract

Survival models with a cure fraction are suitable for settings in which a portion of the population may remain free of the event of interest in the long term, a behavior associated with the presence of a plateau in the survival curve. This dissertation proposes a class of Bayesian semiparametric cumulative hazard transformation models with a cure fraction, in which the incidence component represents the cure probability through a *logit* link, while the latency component describes the time to the event among susceptible individuals. Previously specified transformations were considered, including proportional hazards and proportional odds structures, as well as a data-adaptive monotone transformation estimated from the data through a generator constructed from cubic *B-spline* basis functions. The *baseline* cumulative hazard was represented nonparametrically by positive increments defined at the observed event times, with regularization based on second-order random-walk penalties. Posterior inference was conducted using a Metropolis-within-Gibbs algorithm combining multivariate random-walk proposals, preconditioned MALA updates, and Gibbs steps. The performance of the adaptive model was evaluated through a Monte Carlo simulation study with sample sizes of 30, 50, 100, and 500 and censoring levels of 10% and 30%. The results showed low bias values for the latency coefficients, a reduction in RMSE as the sample size increased, and recovery of the transformation function and the *baseline* cumulative hazard, while the incidence and transformation components exhibited greater uncertainty in some scenarios. The application was conducted using a retrospective cohort of breast cancer patients from the Netherlands Cancer Institute. The adaptive model presented the lowest DIC among the Bayesian models evaluated and a WAIC close to the lowest value obtained by the proportional hazards model. Among susceptible patients, tumor grade III was associated with earlier event occurrence. The prior-sensitivity analysis indicated stability of the regression effects and estimated functions, while the posterior correlation analysis showed dependence between some incidence and latency effects. The results show that the proposed approach allows the risk transformation to be estimated from the data while preserving the separation between the mechanisms associated with cure and the survival of susceptible individuals.

Keywords: *survival analysis, cure fraction, Bayesian inference, transformation models, semiparametric models.*

Lista de Figuras

5.1	Curva de Kaplan–Meier para mulheres diagnosticadas com câncer de mama no conjunto de dados do NKI.	49
5.2	Correlações a posteriori entre os coeficientes do componente de latência (β) e os do componente de incidência (γ), sob a especificação adaptativa. . . .	52
5.3	Estimativa de Kaplan–Meier e curvas ajustadas de sobrevivência marginal para os modelos adaptativo, PH, PO, Weibull e Cox padrão de mistura com fração de cura. As bandas sombreadas representam intervalos de credibilidade posterior de 95% para os ajustes Bayesianos.	53
5.4	Sensibilidade aos valores das distribuições a priori sob hipera-priors de precisão de suavização Gamma(1, 1) e Gamma(0.1, 0.1): médias posteriores e bandas de credibilidade de 95	54
5.5	<i>Traceplots</i> das quatro cadeias para os coeficientes de latência β no modelo de riscos proporcionais (RP).	55
5.6	<i>Traceplots</i> das quatro cadeias para os coeficientes de latência β no modelo de chances proporcionais (CP).	55
5.7	<i>Traceplots</i> das quatro cadeias para os coeficientes de latência β no modelo com transformação adaptativa.	56
5.8	<i>Traceplots</i> das quatro cadeias para os coeficientes de latência β no modelo com transformação adaptativa sob a hiperpriori Gama(0,1; 0,1).	56
5.9	<i>Traceplots</i> das quatro cadeias para os coeficientes de incidência γ no modelo de riscos proporcionais (RP).	57
5.10	<i>Traceplots</i> das quatro cadeias para os coeficientes de incidência γ no modelo de chances proporcionais (CP).	57
5.11	<i>Traceplots</i> das quatro cadeias para os coeficientes de incidência γ no modelo com transformação adaptativa.	58
5.12	<i>Traceplots</i> das quatro cadeias para os coeficientes de incidência γ no modelo com transformação adaptativa sob a hiperpriori Gama(0,1; 0,1).	58
5.13	<i>Traceplots</i> das quatro cadeias para os coeficientes de transformação θ no modelo com transformação adaptativa.	59
5.14	<i>Traceplots</i> das quatro cadeias para os coeficientes de transformação θ no modelo com transformação adaptativa sob a hiperpriori Gama(0,1; 0,1). . .	59

Lista de Tabelas

4.1	Configuração da simulação, dimensão ajustada e desempenho no nível da função. O ESS apresentado corresponde ao valor mínimo entre todas as replicações registradas e os blocos de diagnóstico.	44
4.2	Características operacionais por bloco de parâmetros. São apresentados o viés absoluto máximo, o intervalo de RMSE e o intervalo de cobertura empírica de 95% entre os coeficientes pertencentes a cada bloco.	45
5.1	Características <i>baseline</i> e dados operatórios da coorte retrospectiva analisada por van de Vijver et al. (2002).	48
5.2	Diagnósticos de convergência das quatro cadeias para os principais blocos de coeficientes na aplicação aos dados do NKI.	50
5.3	Resumos a posteriori para o modelo de cura com transformação adaptativa proposto na aplicação do NKI.	51
5.4	Comparação dos modelos. WAIC e DIC utilizam a mesma verossimilhança baseada nos intervalos de ocorrência dos eventos para todos os modelos Bayesianos.	52

Sumário

Resumo	7
Abstract	8
Lista de Figuras	9
Lista de Tabelas	9
Capítulos	10
1 Conceitos Preliminares	13
1.1 Estrutura dos Dados de Sobrevida	15
1.2 Métodos Estatísticos Para Dados de Sobrevida	16
1.2.1 Estimador de Kaplan–Meier	16
1.2.2 Modelo de Riscos Proporcionais de Cox	18
1.2.3 Limitações da Hipótese de Riscos Proporcionais	21
1.3 Sobreviventes de Longo Prazo	23
1.3.1 Modelos de Cura de Mistura	24
1.3.2 Modelos de Cura de Não-Mistura	26
1.4 Organização da Dissertação	27
2 Objetivos	29
2.1 Geral	29
2.2 Específicos	29
3 Framework Bayesiano Para Modelos Semiparamétrico de Transformação do Risco com Fração de Cura	31
3.1 Especificação do Modelo de Transformação	31
3.2 Extensão para Sobrevida de Longo Prazo	32
3.3 Especificação da Priori Penalizada	33
3.4 Inferência a Posteriori e Computação	36
3.5 Especificação de uma Transformação Adaptativa aos Dados	37
3.6 Critérios Bayesianos de Comparação de Modelos	38
4 Estudo de Simulação	41
4.1 Configuração do Estudo de Simulação	41
4.2 Critérios de Avaliação	42
4.3 Resultados do Estudo de Simulação	43
4.3.1 Desempenho Computacional e Recuperação das Funções	43
4.3.2 Desempenho dos Parâmetros	44

5	Aplicação em Dados de Câncer de Mama	46
5.1	Introdução	46
5.2	Dados do Estudo	47
5.3	Especificação e Ajuste dos Modelos	48
5.4	Resultados da Aplicação	50
5.4.1	Diagnósticos de Convergência	50
5.4.2	Inferência para o Modelo Adaptativo	51
5.4.3	Comparação entre Modelos	52
5.4.4	Análise de Sensibilidade e Diagnósticos Gráficos	53
6	Conclusão do trabalho	60
	Referências	61

Conceitos Preliminares

A análise de sobrevivência compreende um conjunto de métodos estatísticos destinados ao estudo do tempo decorrido entre uma origem temporal previamente definida e a ocorrência de um evento de interesse. Embora tenha se desenvolvido principalmente em estudos médicos e epidemiológicos, essa área possui aplicações em diferentes campos do conhecimento, sempre que o desfecho investigado puder ser representado pelo tempo necessário para a ocorrência de determinado evento. Em aplicações médicas, o evento de interesse pode corresponder à morte, à recorrência de uma doença, ao aparecimento de uma complicação, à remissão de uma condição clínica, à resposta a um tratamento ou à progressão de uma enfermidade. Em estudos industriais e de confiabilidade, o evento pode representar a falha de uma máquina, de um componente eletrônico ou de um sistema produtivo. Em outros contextos, pode-se investigar o tempo até a interrupção de um serviço, a duração de um contrato, a reincidência criminal, a dissolução de uma união ou qualquer outro desfecho cuja ocorrência esteja associada a uma escala temporal. Por essa razão, a expressão “tempo de sobrevivência” não deve ser interpretada exclusivamente como o tempo até a morte. Em sentido estatístico, ela designa o tempo até a ocorrência de qualquer evento previamente estabelecido pelo pesquisador. De modo semelhante, os termos “falha” e “evento” são utilizados de forma genérica e não implicam necessariamente um resultado negativo. Em determinados estudos, por exemplo, o evento pode ser a recuperação de um paciente, a obtenção de resposta terapêutica ou a conclusão bem-sucedida de um procedimento (Maller and Zhou, 1996; Klein and Moeschberger, 1997; Collett, 2015; Schober and Vetter, 2018).

A definição adequada do tempo de sobrevivência exige a especificação de três elementos fundamentais: a *origem* da escala temporal, o *evento de interesse* e a *unidade* na qual o tempo será mensurado. A *origem temporal*, também chamada *tempo zero*, deve corresponder a um marco identificável e comum aos indivíduos estudados. Quando a escala considerada é a idade, a origem pode ser o nascimento; em um ensaio clínico, pode corresponder à inclusão do participante ou ao início do tratamento; em estudos oncológicos, pode ser a data do diagnóstico, da cirurgia ou do início de uma terapia; em análises de confiabilidade, pode representar o momento em que o equipamento é colocado em funcionamento (Klein and Moeschberger, 1997; Meeker and Escobar, 1998). A escolha da origem temporal deve ser coerente com o objetivo científico da investigação, visto que origens distintas podem produzir interpretações diferentes para um mesmo desfecho. Em estudos de câncer, por exemplo, o tempo desde o diagnóstico mede uma trajetória diferente daquela obtida quando o acompanhamento é iniciado na data da cirurgia. Por esse motivo, a origem da escala temporal não constitui apenas uma decisão operacional, mas parte essencial da definição do problema estatístico (Li and Ma, 2013).

A *unidade de tempo* também depende do contexto da aplicação. Segundo [Li and Ma \(2013\)](#) e [Collett \(2015\)](#), o acompanhamento pode ser expresso em dias, meses, anos, ciclos de funcionamento ou qualquer outra medida adequada ao fenômeno investigado. Independentemente da unidade utilizada, espera-se que os tempos observados sejam não negativos e que a escala seja suficientemente precisa para representar a ocorrência dos eventos. Em algumas aplicações, diferentes indivíduos podem apresentar o evento no mesmo instante registrado, produzindo empates entre os tempos observados, aspecto que deve ser considerado na formulação e na implementação dos métodos estatísticos.

Os dados utilizados em análise de sobrevivência podem ser obtidos por meio de estudos *retrospectivos* ou *prospectivos*. Nos *estudos retrospectivos*, as informações sobre exposição, acompanhamento e ocorrência do evento são recuperadas a partir de registros previamente existentes, como prontuários médicos, sistemas hospitalares, bancos administrativos ou bases de vigilância. Nesse contexto, uma coorte histórica pode ser reconstruída utilizando indivíduos que iniciaram o acompanhamento em um período anterior ao início formal da pesquisa. Já nos *estudos prospectivos*, os indivíduos são acompanhados a partir de um momento previamente definido, e as informações são registradas à medida que o acompanhamento ocorre. Esses estudos podem assumir a forma de *coortes observacionais* ou *ensaios clínicos randomizados* ([Lilienfeld and Lilienfeld, 1980](#); [Hartung and Touchette, 2009](#); [Wang and Kattan, 2020](#); [Kiani et al., 2022](#)).

Nas *coortes observacionais*, busca-se descrever a ocorrência do evento e investigar fatores associados à evolução dos indivíduos em condições naturais de exposição. Esses estudos podem contribuir para a caracterização quantitativa da história natural de uma doença, para a identificação de fatores prognósticos e para o planejamento de serviços clínicos e de políticas de saúde pública ([Wang and Kattan, 2020](#)). Nos *ensaios clínicos randomizados*, por sua vez, os participantes são alocados aleatoriamente a diferentes intervenções, permitindo avaliar os efeitos de tratamentos, procedimentos ou estratégias de prevenção sobre o tempo até a ocorrência de determinado desfecho. Nessas situações, a análise de sobrevivência possibilita comparar não apenas a proporção de indivíduos que experimentam o evento, mas também o momento em que ele ocorre durante o período de acompanhamento. A escolha entre uma coorte observacional, um ensaio clínico ou um estudo retrospectivo depende dos objetivos da pesquisa, da natureza da doença ou do fenômeno investigado, de aspectos éticos e das limitações práticas relacionadas à coleta dos dados ([Chow and Chang, 2008](#); [Zabor et al., 2020](#)).

Embora os indivíduos sejam acompanhados ao longo do tempo, *os dados de sobrevivência não são necessariamente longitudinais* no sentido estrito de envolverem medidas repetidas de um mesmo desfecho. Em sua forma mais simples, cada participante contribui com um único tempo até o evento, acompanhado de informações sobre sua ocorrência e de um conjunto de covariáveis. Estudos de sobrevivência podem, entretanto, ser combinados com dados longitudinais quando características clínicas, biomarcadores ou outras variáveis são observadas repetidamente durante o acompanhamento. Outra particularidade importante é que o período de observação pode variar entre os indivíduos. Alguns participantes podem ingressar no estudo em momentos distintos, ser acompanhados por durações diferentes ou deixar de ser observados antes da ocorrência do evento. Consequentemente, nem todos os tempos de sobrevivência são conhecidos de forma completa. A presença dessas observações incompletas, denominadas *censuradas*, constitui uma das características principais dos dados de sobrevivência e exige a utilização de métodos estatísticos desenvolvidos para incorporar a informação parcial disponível ([Klein and Moeschberger, 1997](#); [Leboo, 2023](#)).

1.1 Estrutura dos Dados de Sobrevivência

Uma das principais características dos dados de sobrevivência é a possibilidade de o evento de interesse não ser observado para todos os indivíduos durante o período de acompanhamento. Nesses casos, embora o tempo exato até o evento seja desconhecido, parte da informação sobre esse tempo permanece disponível. Essa observação incompleta é denominada *censura* e constitui uma característica estrutural dos dados de sobrevivência, não devendo ser tratada simplesmente como uma informação ausente. A censura ocorre quando o tempo até o evento não pode ser determinado exatamente, mas é possível estabelecer alguma informação parcial sobre sua localização na escala temporal. Dependendo do delineamento do estudo e da forma como o acompanhamento é realizado, essa informação pode indicar que o evento ocorreu após determinado instante, antes de determinado instante ou dentro de um intervalo conhecido (Klein and Moeschberger, 1997; Giolo and Colosimo, 2006; Collett, 2015).

A forma mais comum em aplicações médicas e epidemiológicas é a *censura à direita*. Ela ocorre quando o acompanhamento de um indivíduo é encerrado antes que o evento de interesse seja observado, de modo que se sabe apenas que o verdadeiro tempo até o evento é superior ao último tempo registrado. Essa situação pode ocorrer porque o estudo atingiu sua data administrativa de encerramento, porque o indivíduo deixou de comparecer às avaliações, porque houve perda de contato ou porque o acompanhamento foi interrompido por alguma outra razão. A *censura administrativa* é um caso particular de censura à direita e ocorre quando o estudo termina em uma data previamente estabelecida. Os indivíduos que ainda não apresentaram o evento até esse momento são considerados censurados em seus respectivos tempos de acompanhamento. Esse mecanismo é frequente em estudos clínicos e epidemiológicos, uma vez que limitações financeiras, operacionais e éticas geralmente impedem que todos os participantes sejam acompanhados até a ocorrência do evento. Outro mecanismo de censura à direita decorre da *perda de acompanhamento*. Nessa situação, o indivíduo deixa de ser observado antes do término previsto do estudo, e seu estado em relação ao evento passa a ser desconhecido. A perda de acompanhamento deve ser investigada cuidadosamente, pois pode introduzir viés quando sua ocorrência está associada ao prognóstico ou ao verdadeiro tempo até o evento (Klein and Moeschberger, 1997; Giolo and Colosimo, 2006; Li and Ma, 2013; Collett, 2015; Leboo, 2023).

Para definir matematicamente a censura à direita, considere T_i como o verdadeiro tempo até a ocorrência do evento de interesse para o indivíduo i e C_i como o correspondente tempo de censura à direita. O tempo observado, denotado por Y_i , e o indicador de ocorrência do evento, denotado por δ_i , são definidos por:

$$Y_i = \min(T_i, C_i), \quad \delta_i = \mathbb{I}(T_i \leq C_i) \quad (1.1)$$

em que $\mathbb{I}(\cdot)$ representa a *função indicadora*. Quando $\delta_i = 1$, o evento foi observado no tempo $Y_i = T_i$. Quando $\delta_i = 0$, o indivíduo foi censurado no tempo $Y_i = C_i$, sendo possível afirmar apenas que $T_i > Y_i$. A representação por meio do par (Y_i, δ_i) é fundamental para a construção da função de verossimilhança em modelos de sobrevivência. Assim, para os indivíduos que apresentam o evento, a contribuição para a verossimilhança envolve a função densidade do tempo de sobrevivência, e, para os indivíduos censurados à direita, a contribuição corresponde à probabilidade de sobrevivência além do tempo observado. Dessa forma, mesmo sem conhecer o tempo exato até o evento, a observação censurada fornece informação relevante para a estimação do modelo (Klein and Moeschberger, 1997; Collett, 2015).

Além do mecanismo de censura à direita, pode ocorrer *censura à esquerda*. Nesse caso, sabe-se que o evento ocorreu antes do primeiro instante em que o indivíduo foi avaliado, mas o tempo exato de sua ocorrência é desconhecido. Se L_i representa o primeiro tempo de avaliação, a informação disponível pode ser expressa por $T_i \leq L_i$. Esse mecanismo pode surgir, por exemplo, em estudos ou situações nos quais uma doença ou condição clínica já estava presente na primeira avaliação, sem que seja possível determinar exatamente quando ela se iniciou (Klein and Moeschberger, 1997; Lee and Wang, 2003).

Outra característica possível na estrutura dos dados de sobrevivência é o *truncamento à esquerda*, também denominado *entrada tardia*. Nesse caso, o indivíduo somente é incluído no conjunto observado se sobreviver ou permanecer livre do evento até o instante de entrada no estudo. Se E_i representa o tempo de entrada, a observação do indivíduo está condicionada a $T_i \geq E_i$. Diferentemente da censura, na qual o indivíduo pertence à população observada, mas fornece informação incompleta sobre o tempo até o evento, no truncamento os indivíduos que apresentam o evento antes do instante de entrada não são observados (Klein and Moeschberger, 1997).

Os mecanismos de censura e de truncamento devem ser considerados tanto na escolha do método estatístico quanto na interpretação dos resultados. Métodos que ignoram essas características podem produzir estimativas enviesadas da sobrevivência, do risco e dos efeitos das covariáveis. Por esse motivo, foram desenvolvidos procedimentos específicos para utilizar adequadamente a informação parcial contida nas observações censuradas. Entre esses procedimentos, destacam-se os estimadores não paramétricos da função de sobrevivência, como o estimador de Kaplan-Meier (Kaplan and Meier, 1958), os testes para comparação de grupos (Bradburn et al., 2003) e os modelos de regressão para o tempo até o evento, como o modelo de *risco proporcional de Cox* (Cox, 1972). Essas abordagens são apresentadas na próxima seção, com ênfase na incorporação de covariáveis e na modelagem do risco de ocorrência do evento.

1.2 Métodos Estatísticos Para Dados de Sobrevivência

1.2.1 Estimador de Kaplan–Meier

A análise inicial de dados de sobrevivência é frequentemente realizada por meio de métodos não paramétricos, os quais permitem estimar características da distribuição do tempo até o evento sem a necessidade de especificar uma distribuição de probabilidade paramétrica. Entre esses procedimentos, destaca-se o estimador de Kaplan–Meier, também denominado estimador produto-limite, utilizado para estimar a função de sobrevivência na presença de observações censuradas à direita (Kaplan and Meier, 1958). Para defini-lo, seja T uma variável aleatória positiva que representa o tempo até a ocorrência do evento de interesse. A função de sobrevivência é definida por:

$$S(t) = P(T > t), \quad t \geq 0 \quad (1.2)$$

A quantidade $S(t)$ representa a probabilidade de um indivíduo permanecer livre do evento além do tempo t . Por definição, a função de sobrevivência é não crescente, contínua à direita e satisfaz:

$$S(0) = 1, \quad \lim_{t \rightarrow \infty} S(t) = 0 \quad (1.3)$$

A última condição corresponde a uma distribuição de sobrevivência própria, na qual todos os indivíduos eventualmente apresentam o evento. Em populações com sobreviventes de longo prazo, entretanto, pode ocorrer:

$$\lim_{t \rightarrow \infty} S(t) = \rho > 0 \quad (1.4)$$

em que ρ representa a proporção de indivíduos que não experimentarão o evento em longo prazo. Esse comportamento pode constituir um indício empírico da presença de uma possível fração de cura, desde que o período de acompanhamento e o número de indivíduos sob risco nos tempos finais sejam adequados.

Agora, considere uma amostra composta por n indivíduos e sejam $t_{(1)} < \dots < t_{(J)}$ os tempos distintos nos quais ocorreram eventos. Para cada instante $t_{(j)}$, denote por d_j o número de eventos observados e por n_j o número de indivíduos sob risco imediatamente antes de $t_{(j)}$. A probabilidade condicional de sobreviver ao instante $t_{(j)}$, dado que o indivíduo permaneceu sob risco até esse momento, é estimada por:

$$1 - \frac{d_j}{n_j} = \frac{n_j - d_j}{n_j} \quad (1.5)$$

O *estimador de Kaplan–Meier* (Kaplan and Meier, 1958) da função de sobrevivência é, então, obtido pelo produto dessas probabilidades condicionais ao longo dos tempos de evento, isto é,

$$\hat{S}(t) = \prod_{j: t_{(j)} \leq t} \left(1 - \frac{d_j}{n_j} \right) \quad (1.6)$$

Para tempos anteriores ao primeiro evento observado, tem-se $\hat{S}(t) = 1$. A cada tempo de evento, a curva apresenta uma redução cuja magnitude depende da proporção de eventos entre os indivíduos que permaneciam sob risco naquele instante. Entre dois tempos consecutivos de evento, a estimativa permanece constante, de modo que $\hat{S}(t)$ é uma função em degraus, não crescente e contínua à direita. As observações censuradas não provocam reduções diretas na curva de sobrevivência, pois não correspondem à ocorrência do evento. Entretanto, um indivíduo censurado deixa de pertencer ao conjunto de risco após seu tempo de censura, reduzindo os valores de n_j nos tempos de evento subsequentes. Dessa forma, as observações censuradas contribuem para a estimação até o último instante em que o indivíduo permaneceu sob acompanhamento (Collett, 2015).

A variabilidade do estimador de Kaplan–Meier pode ser estimada por meio da fórmula de Greenwood (Greenwood, 1926). Neste caso, para um tempo t , tem-se que a variância estimada de $\hat{S}(t)$ é dada por:

$$\widehat{\text{Var}} \{ \hat{S}(t) \} = \hat{S}(t)^2 \sum_{j: t_{(j)} \leq t} \frac{d_j}{n_j(n_j - d_j)} \quad (1.7)$$

e o erro-padrão correspondente é definido por:

$$\widehat{\text{EP}} \{ \hat{S}(t) \} = \sqrt{\widehat{\text{Var}} \{ \hat{S}(t) \}} \quad (1.8)$$

A construção direta de intervalos de confiança simétricos em torno de $\hat{S}(t)$ pode produzir limites inferiores menores que zero ou limites superiores maiores que um. Por esse motivo, transformações que preservam o intervalo $[0, 1]$, como a transformação logarítmica complementar, são frequentemente empregadas. Um intervalo de confiança aproximado de nível $100(1 - \alpha)\%$ para $S(t)$ pode ser obtido na escala $\log\{-\log S(t)\}$ e posteriormente transformado para a escala original (Collett, 2015).

As curvas de Kaplan–Meier podem ser estimadas separadamente para diferentes grupos, permitindo uma análise descritiva das diferenças entre suas distribuições de sobrevivência. A comparação deve considerar não apenas a posição global das curvas, mas também a possível ocorrência de cruzamentos, a quantidade de indivíduos sob risco ao longo do acompanhamento, a proporção de censura e a incerteza associada às estimativas. A igualdade entre as funções de sobrevivência de dois ou mais grupos pode ser avaliada por meio de testes não paramétricos baseados em postos. Entre eles, o *teste log-rank* é um dos mais utilizados e compara, em cada tempo de evento, o número observado de eventos em cada grupo com o número esperado sob a hipótese de igualdade das funções de sobrevivência (Bradburn et al., 2003). Para defini-lo, considere G grupos e, para cada tempo de evento $t_{(j)}$, sejam d_{gj} e n_{gj} , respectivamente, o número de eventos e o número de indivíduos sob risco no grupo g . Se d_j e n_j representam os valores totais entre os grupos, o número esperado de eventos no grupo g , sob a hipótese nula, é dado por:

$$e_{gj} = \frac{n_{gj}d_j}{n_j} \quad (1.9)$$

A estatística do teste é construída a partir das diferenças acumuladas entre os números observados e esperados de eventos. Sob a hipótese nula de igualdade das funções de sobrevivência, essa estatística apresenta, assintoticamente, distribuição qui-quadrado com $G - 1$ graus de liberdade. O teste log-rank atribui pesos semelhantes aos diferentes tempos de evento e apresenta maior sensibilidade quando as razões de riscos entre os grupos são aproximadamente constantes. Quando as curvas de sobrevivência se cruzam ou os efeitos variam ao longo do tempo, o teste pode apresentar baixa capacidade de detectar diferenças, sendo necessário considerar procedimentos alternativos ou modelos que permitam efeitos dependentes do tempo (Bradburn et al., 2003; Collett, 2015).

1.2.2 Modelo de Riscos Proporcionais de Cox

Embora o estimador de Kaplan–Meier seja adequado para descrever a distribuição dos tempos de sobrevivência e comparar grupos definidos por uma única característica, sua utilização torna-se limitada quando se deseja avaliar simultaneamente a associação entre o tempo até o evento e múltiplas covariáveis. Nessa situação, são necessários modelos de regressão capazes de incorporar a censura e quantificar os efeitos ajustados das características dos indivíduos. Entre as abordagens mais utilizadas encontra-se o modelo de riscos proporcionais proposto por Cox (1972). Esse modelo relaciona a função de risco a um conjunto de covariáveis sem exigir a especificação de uma distribuição paramétrica para o tempo de sobrevivência ou para a função *baseline* de risco. Por essa razão, o modelo de Cox é classificado como semiparamétrico: os efeitos das covariáveis são representados por um componente paramétrico, enquanto a função *baseline* de risco permanece não especificada (Klein and Moeschberger, 1997; Collett, 2015).

Seja T uma variável aleatória não negativa que representa o tempo até a ocorrência do evento de interesse. A função de risco no tempo t é definida por:

$$h(t) = \lim_{\Delta t \rightarrow 0^+} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \quad (1.10)$$

A função $h(t)$ representa a taxa instantânea de ocorrência do evento no tempo t , condicionada ao fato de o indivíduo ter permanecido livre do evento até esse instante. Embora seja frequentemente denominada risco, essa função não representa uma probabilidade e, portanto, não está restrita ao intervalo $[0, 1]$. Quando T possui função densidade $f(t)$ e função de sobrevivência $S(t)$, a função de risco pode ser expressa por:

$$h(t) = \frac{f(t)}{S(t)} \quad (1.11)$$

A função de risco acumulado, em particular, é definida como a integral da função de risco no intervalo entre zero e t , isto é,

$$\Lambda(t) = \int_0^t h(u) du \quad (1.12)$$

e sua relação com a função de sobrevivência é dada por:

$$S(t) = \exp\{-\Lambda(t)\} \quad (1.13)$$

A partir dessas relações, considere $\mathbf{X} = (X_1, \dots, X_p)^\top$ como um vetor de covariáveis associado a determinado indivíduo e o vetor de coeficientes de regressão correspondente, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$. Segundo [Cox \(1972\)](#), a *função de risco condicional*, neste caso, é descrita por:

$$h(t \mid \mathbf{X}) = h_0(t) \exp(\mathbf{X}^\top \boldsymbol{\beta}) \quad (1.14)$$

em que $h_0(t)$ é a função *baseline* de risco. Essa função representa o risco no tempo t para um indivíduo cujo vetor de covariáveis é $\mathbf{X} = \mathbf{0}$. Quando o valor zero não possui interpretação natural para alguma covariável, pode-se centralizá-la em torno de um valor de referência clinicamente relevante, de modo a tornar a interpretação do risco *baseline* mais apropriada. O termo $\exp(\mathbf{X}^\top \boldsymbol{\beta})$ representa o efeito multiplicativo das covariáveis sobre a função *baseline* de risco. Consequentemente, a função de risco cumulativo condicional é dada por:

$$\Lambda(t \mid \mathbf{X}) = \Lambda_0(t) \exp(\mathbf{X}^\top \boldsymbol{\beta}) \quad (1.15)$$

em que

$$\Lambda_0(t) = \int_0^t h_0(u) du \quad (1.16)$$

Uma das principais características do modelo de Cox é a possibilidade de estimar $\boldsymbol{\beta}$ sem especificar parametricamente a função *baseline* de risco. Essa estimação é realizada por meio da *verossimilhança parcial* proposta por [Cox \(1972\)](#). Considere uma amostra

composta por n indivíduos, para os quais são observados os pares (Y_i, δ_i) e os vetores de covariáveis $\mathbf{X}_i$, com $i = 1, \dots, n$. Seja $\mathcal{R}(Y_i)$ o conjunto de indivíduos sob risco imediatamente antes do tempo Y_i , definido por:

$$\mathcal{R}(Y_i) = \{j : Y_j \geq Y_i\} \quad (1.17)$$

Na ausência de empates entre os tempos de evento, a probabilidade condicional de que o indivíduo i apresente o evento no tempo Y_i , dado que um evento ocorreu entre os indivíduos pertencentes a $\mathcal{R}(Y_i)$, é:

$$\frac{\exp(\mathbf{X}_i^\top \boldsymbol{\beta})}{\sum_{j \in \mathcal{R}(Y_i)} \exp(\mathbf{X}_j^\top \boldsymbol{\beta})} \quad (1.18)$$

A *função de verossimilhança parcial* é, então, obtida pelo produto dessas contribuições sobre todos os indivíduos que apresentaram o evento, isto é,

$$L_p(\boldsymbol{\beta}) = \prod_{i: \delta_i=1} \frac{\exp(\mathbf{X}_i^\top \boldsymbol{\beta})}{\sum_{j \in \mathcal{R}(Y_i)} \exp(\mathbf{X}_j^\top \boldsymbol{\beta})} \quad (1.19)$$

A correspondente função log-verossimilhança parcial é dada por:

$$\ell_p(\boldsymbol{\beta}) = \sum_{i=1}^n \delta_i \left[\mathbf{X}_i^\top \boldsymbol{\beta} - \log \left\{ \sum_{j \in \mathcal{R}(Y_i)} \exp(\mathbf{X}_j^\top \boldsymbol{\beta}) \right\} \right] \quad (1.20)$$

e o *estimador de máxima verossimilhança parcial* de $\boldsymbol{\beta}$ é dado por:

$$\hat{\boldsymbol{\beta}} = \arg \max_{\boldsymbol{\beta}} \ell_p(\boldsymbol{\beta}) \quad (1.21)$$

Após a estimação de $\boldsymbol{\beta}$, a função de risco acumulado *baseline* pode ser estimada pelo estimador de Breslow. Se $t_{(1)} < \dots < t_{(J)}$ representam os tempos distintos de evento e d_j é o número de eventos observados em $t_{(j)}$, o estimador é dado por:

$$\hat{\Lambda}_0(t) = \sum_{j: t_{(j)} \leq t} \frac{d_j}{\sum_{i \in \mathcal{R}(t_{(j)})} \exp(\mathbf{X}_i^\top \hat{\boldsymbol{\beta}})} \quad (1.22)$$

A função de sobrevivência *baseline* pode ser estimada por:

$$\hat{S}_0(t) = \exp \left\{ -\hat{\Lambda}_0(t) \right\} \quad (1.23)$$

e, assim, para um indivíduo com vetor de covariáveis $\mathbf{X}$, a função de sobrevivência ajustada é estimada por:

$$\hat{S}(t | \mathbf{X}) = \exp \left\{ -\hat{\Lambda}_0(t) \exp(\mathbf{X}^\top \hat{\boldsymbol{\beta}}) \right\}. \quad (1.24)$$

Além disso, a inferência sobre os coeficientes β pode ser realizada por meio de testes de Wald, da razão de verossimilhanças ou do escore. Intervalos de confiança aproximados para β_j podem ser construídos a partir da normalidade assintótica de $\hat{\beta}_j$. Para a razão de riscos, os limites são obtidos pela exponenciação dos limites correspondentes na escala do coeficiente. Assim, um *intervalo de confiança aproximado* de nível $100(1 - \alpha)\%$ para $\exp(\beta_j)$ é dado por:

$$\left[\exp \left\{ \hat{\beta}_j - z_{1-\alpha/2} \widehat{\text{EP}}(\hat{\beta}_j) \right\}, \exp \left\{ \hat{\beta}_j + z_{1-\alpha/2} \widehat{\text{EP}}(\hat{\beta}_j) \right\} \right] \quad (1.25)$$

em que $z_{1-\alpha/2}$ representa o quantil de ordem $1 - \alpha/2$ da distribuição normal padrão.

1.2.3 Limitações da Hipótese de Riscos Proporcionais

A flexibilidade proporcionada pela especificação não paramétrica da função *baseline* de risco contribui para a ampla utilização do modelo de Cox em estudos com dados de sobrevivência. Entretanto, essa flexibilidade não elimina as restrições impostas à relação entre as covariáveis e o risco de ocorrência do evento. Em particular, o modelo pressupõe que os efeitos das covariáveis sejam multiplicativos e permaneçam constantes durante todo o período de acompanhamento. De fato, considere dois indivíduos com vetores de covariáveis $\mathbf{X}_a$ e $\mathbf{X}_b$. A *razão entre suas funções de risco* é dada por:

$$\frac{h(t \mid \mathbf{X}_a)}{h(t \mid \mathbf{X}_b)} = \exp \{ (\mathbf{X}_a - \mathbf{X}_b)^\top \beta \} \quad (1.26)$$

Como essa razão não depende de t , o modelo pressupõe que os riscos sejam proporcionais ao longo do período de acompanhamento. Isso significa que as funções de risco de dois indivíduos podem variar com o tempo, mas a razão entre elas permanece constante. Para interpretar individualmente o efeito de uma covariável X_j , considere dois indivíduos que diferem em uma unidade nessa covariável e apresentam os mesmos valores nas demais características. A razão de riscos entre eles é $\text{HR}_j = \exp(\beta_j)$, em que HR_j representa a razão de riscos associada a um acréscimo de uma unidade em X_j , mantendo-se constantes as demais covariáveis. Assim:

- Quando $\beta_j > 0$, tem-se $\exp(\beta_j) > 1$, indicando que o aumento da covariável está associado a uma elevação na taxa instantânea de ocorrência do evento.
- Quando $\beta_j < 0$, tem-se $\exp(\beta_j) < 1$, indicando uma redução nessa taxa. Se $\beta_j = 0$, então $\exp(\beta_j) = 1$, o que representa ausência de associação entre a covariável e o risco, condicionada às demais variáveis incluídas no modelo.

Essa proporcionalidade também pode ser expressa em termos da função de sobrevivência. De fato, como:

$$S(t \mid \mathbf{X}) = \exp \{ -\Lambda_0(t) \exp(\mathbf{X}^\top \beta) \} \quad (1.27)$$

segue que:

$$\log [-\log \{S(t \mid \mathbf{X})\}] = \log \{ \Lambda_0(t) \} + \mathbf{X}^\top \beta \quad (1.28)$$

Assim, para dois perfis de covariáveis, a diferença entre as funções transformadas é *constante ao longo do tempo*, isto é,

$$\log [-\log \{S(t \mid \mathbf{X}_a)\}] - \log [-\log \{S(t \mid \mathbf{X}_b)\}] = (\mathbf{X}_a - \mathbf{X}_b)^\top \boldsymbol{\beta} \quad (1.29)$$

Essa relação fornece uma interpretação gráfica da hipótese de riscos proporcionais: curvas de $\log[-\log\{\hat{S}(t)\}]$ construídas para diferentes grupos devem apresentar trajetórias aproximadamente paralelas quando a suposição é razoável. A ausência de paralelismo pode indicar que a razão de riscos se modifica ao longo do tempo. Em muitas aplicações, entretanto, o efeito de uma covariável não permanece constante durante todo o acompanhamento. O benefício de determinado tratamento pode ser mais pronunciado nos primeiros meses e diminuir posteriormente, enquanto o efeito de um fator prognóstico pode se intensificar à medida que a doença evolui. Nesses casos, a representação do efeito por um único coeficiente β_j pode ser insuficiente (Klein and Moeschberger, 1997; Collett, 2015). Uma extensão do modelo permite que os coeficientes dependam do tempo, de modo que:

$$h(t \mid \mathbf{X}) = h_0(t) \exp \left\{ \sum_{j=1}^p \beta_j(t) X_j \right\} \quad (1.30)$$

Assim, para a covariável X_j , a razão de riscos passa a ser dada por:

$$\text{HR}_j(t) = \exp \{ \beta_j(t) \} \quad (1.31)$$

Quando $\beta_j(t)$ varia com t , a razão de riscos também se modifica ao longo do acompanhamento, caracterizando uma violação da hipótese de proporcionalidade. Um valor positivo de $\beta_j(t)$ em determinados períodos e negativo em outros pode indicar, inclusive, uma inversão temporal na direção da associação entre a covariável e o risco. Nessas circunstâncias, o coeficiente constante estimado pelo modelo de Cox pode ser interpretado como uma síntese ponderada dos efeitos observados ao longo dos diferentes tempos de evento. Embora essa medida possa ser útil em algumas aplicações, ela pode ocultar alterações importantes na magnitude ou na direção da associação e dificultar a interpretação clínica dos resultados.

Uma das abordagens mais utilizadas para avaliar a proporcionalidade baseia-se nos resíduos de Schoenfeld (Schoenfeld, 1982). Esses resíduos são definidos nos tempos em que os eventos são observados e representam, para cada covariável, a diferença entre o valor observado para o indivíduo que apresentou o evento e uma média ponderada dos valores daquela covariável entre os indivíduos pertencentes ao conjunto de risco. Sob a hipótese de riscos proporcionais, os resíduos de Schoenfeld escalonados não devem apresentar associação sistemática com o tempo. Em uma representação gráfica, espera-se que a curva suavizada desses resíduos permaneça aproximadamente horizontal. Uma tendência crescente, decrescente ou não linear pode indicar que o efeito da covariável varia ao longo do acompanhamento.

Além da hipótese de riscos proporcionais, o modelo de Cox pressupõe que as covariáveis contínuas atuem linearmente sobre o logaritmo da função de risco. Para uma covariável X_j , a contribuição ao preditor linear é representada por $\beta_j X_j$. Entretanto, em muitas aplicações, a associação entre uma característica contínua e o risco pode apresentar curvatura, limiares, saturação ou mudanças de direção. Essas duas características podem

ser representadas conjuntamente por uma função $f_j(X_j, t)$, de modo que:

$$h(t | \mathbf{X}) = h_0(t) \exp \left\{ \sum_{j=1}^p f_j(X_j, t) \right\} \quad (1.32)$$

Quando $f_j(X_j, t) = f_j(X_j)$, o efeito pode ser não linear, mas permanece constante no tempo. Quando a função também depende de t , são permitidas simultaneamente não linearidade e não proporcionalidade. Algumas violações podem ser tratadas por extensões do próprio modelo de Cox. Efeitos dependentes do tempo podem ser representados por interações entre as covariáveis e funções do tempo. Relações não lineares podem ser incorporadas por transformações, polinômios fracionários ou funções spline. Também é possível utilizar modelos estratificados quando a função *baseline* de risco deve variar entre determinados grupos.

1.3 Sobreviventes de Longo Prazo

Os modelos clássicos de análise de sobrevivência geralmente pressupõem que todos os indivíduos permanecem suscetíveis ao evento de interesse e que, caso o acompanhamento seja suficientemente longo, a probabilidade de sobrevivência tende a zero. Em termos matemáticos, essa condição pode ser expressa por:

$$\lim_{t \rightarrow \infty} S(t | \mathbf{X}) = 0 \quad (1.33)$$

Essa suposição é adequada em situações nas quais todos os indivíduos eventualmente experimentarão o evento, ainda que ele possa ocorrer após um período prolongado. Em algumas aplicações, entretanto, uma proporção não desprezível da população pode permanecer livre do evento mesmo após um horizonte temporal extenso. Esses indivíduos são denominados sobreviventes de longo prazo e sua presença pode tornar inadequados os modelos convencionais que pressupõem a ocorrência eventual do evento para toda a população (Maller and Zhou, 1996; Othus et al., 2012). Essa situação é frequente em estudos oncológicos nos quais o evento corresponde à recorrência da doença, à progressão tumoral ou à morte por uma causa específica. Após determinado período sem a ocorrência do evento, alguns pacientes podem apresentar risco suficientemente reduzido para serem considerados estatisticamente curados. Estruturas semelhantes também podem surgir em estudos de transplantes, doenças cardiovasculares, infecções, fertilidade, reincidência criminal, inadimplência e confiabilidade de sistemas.

O termo “cura” deve ser interpretado com cautela. Em um modelo estatístico, a cura não implica necessariamente a eliminação biológica da doença ou a impossibilidade absoluta de ocorrência futura do evento. A fração de cura representa uma subpopulação que, em relação ao evento definido e dentro do horizonte considerado, não permanece suscetível à sua ocorrência. Dependendo da aplicação, pode ser mais apropriado empregar expressões como “livre do evento em longo prazo”, “não suscetível” ou “sobrevivente de longo prazo” (Othus et al., 2012). A presença de sobreviventes de longo prazo pode ser representada por uma função de sobrevivência imprópria, cuja assíntota é estritamente positiva. Nesse caso,

$$\lim_{t \rightarrow \infty} S(t | \mathbf{X}, \mathbf{Z}) = \rho(\mathbf{Z}), \quad 0 < \rho(\mathbf{Z}) < 1 \quad (1.34)$$

em que $\rho(\mathbf{Z})$ representa a probabilidade de um indivíduo com vetor de covariáveis $\mathbf{Z}$ pertencer à subpopulação curada ou livre do evento em longo prazo.

A distinção entre indivíduos curados e indivíduos suscetíveis censurados é particularmente difícil, pois o estado de cura não é diretamente observado. Para um indivíduo que apresenta o evento, sabe-se que ele pertence à subpopulação suscetível. Para um indivíduo censurado, contudo, não é possível determinar diretamente se ele é curado ou se permanece suscetível, mas não foi acompanhado por tempo suficiente para apresentar o evento. Assim, a estimação da fração de cura depende da separação entre duas fontes de heterogeneidade: a heterogeneidade associada à suscetibilidade ao evento e a heterogeneidade relacionada ao tempo de ocorrência entre os indivíduos suscetíveis. Esses dois mecanismos são usualmente denominados componentes de *incidência* e de *latência*, respectivamente. O componente de *incidência* descreve a probabilidade de pertencimento às subpopulações curada e suscetível. O componente de *latência* descreve a distribuição do tempo até o evento entre os indivíduos suscetíveis. As covariáveis incluídas nesses componentes não precisam ser necessariamente as mesmas, pois uma característica pode influenciar a probabilidade de cura, o tempo até o evento entre os suscetíveis ou ambos (Maller and Zhou, 1996; Tsodikov et al., 2003; Vahidpour, 2016).

1.3.1 Modelos de Cura de Mistura

Os modelos de cura de mistura representam a população como uma combinação de duas subpopulações latentes: uma composta por indivíduos suscetíveis ao evento e outra formada por indivíduos curados ou livres do evento em longo prazo. Essa estrutura permite separar diretamente a probabilidade de cura da distribuição do tempo até o evento entre os indivíduos suscetíveis (Maller and Zhou, 1996).

Considere U_i como uma variável latente indicadora da suscetibilidade do indivíduo i , definida por:

$$U_i = \begin{cases} 1, & \text{se o indivíduo é suscetível} \\ 0, & \text{se o indivíduo é curado} \end{cases} \quad (1.35)$$

Seja $\rho(\mathbf{Z}_i)$ a probabilidade de cura condicionada ao vetor de covariáveis $\mathbf{Z}_i$. Assim,

$$\begin{aligned} P(U_i = 0 \mid \mathbf{Z}_i) &= \rho(\mathbf{Z}_i) \\ P(U_i = 1 \mid \mathbf{Z}_i) &= 1 - \rho(\mathbf{Z}_i) \end{aligned} \quad (1.36)$$

Denote por $S_u(t \mid \mathbf{X}_i)$ a função de sobrevivência condicional entre os indivíduos suscetíveis, em que $\mathbf{X}_i$ representa o vetor de covariáveis associado ao componente de latência. Para um indivíduo curado, o tempo até o evento é considerado infinito e, portanto, sua função de sobrevivência é igual a um para qualquer tempo finito. A função de sobrevivência populacional é obtida pela média das funções de sobrevivência das duas subpopulações,

$$S(t \mid \mathbf{X}_i, \mathbf{Z}_i) = \rho(\mathbf{Z}_i) + \{1 - \rho(\mathbf{Z}_i)\} S_u(t \mid \mathbf{X}_i) \quad (1.37)$$

Essa expressão caracteriza o *modelo clássico de cura de mistura*. A primeira parcela corresponde à contribuição dos indivíduos curados, enquanto a segunda representa a sobrevivência dos indivíduos suscetíveis ponderada pela probabilidade de pertencimento a

essa subpopulação. Se a função de sobrevivência dos suscetíveis for própria, de modo que:

$$\lim_{t \rightarrow \infty} S_u(t | \mathbf{X}_i) = 0 \quad (1.38)$$

então a função de sobrevivência populacional satisfaz:

$$\lim_{t \rightarrow \infty} S(t | \mathbf{X}_i, \mathbf{Z}_i) = \rho(\mathbf{Z}_i) \quad (1.39)$$

Portanto, a assíntota da função de sobrevivência populacional corresponde à probabilidade de cura. Além disso, a probabilidade de cura pode ser relacionada às covariáveis por meio de uma função de ligação. Uma escolha frequente é a ligação logística,

$$\rho(\mathbf{Z}_i) = \frac{\exp(\mathbf{Z}_i^\top \boldsymbol{\zeta})}{1 + \exp(\mathbf{Z}_i^\top \boldsymbol{\zeta})} \quad (1.40)$$

em que $\boldsymbol{\zeta}$ representa o vetor de coeficientes do componente de incidência. Para uma covariável Z_{ij} , a quantidade $\exp(\zeta_j)$ representa a razão de chances de cura associada ao aumento de uma unidade nessa covariável. Assim, um valor positivo de ζ_j está associado ao aumento da probabilidade de cura, enquanto um valor negativo está associado à sua redução.

Por outro lado, entre os indivíduos suscetíveis, a função de sobrevivência $S_u(t | \mathbf{X}_i)$ pode ser especificada por modelos paramétricos, semiparamétricos ou não paramétricos. Dessa forma, o componente de latência pode ser representado por um modelo de riscos proporcionais, de chances proporcionais, de tempo de falha acelerada ou por uma estrutura de transformação mais flexível. Se $f_u(t | \mathbf{X}_i)$ denota a função densidade do tempo até o evento entre os suscetíveis, a densidade populacional para tempos finitos é dada por:

$$f(t | \mathbf{X}_i, \mathbf{Z}_i) = \{1 - \rho(\mathbf{Z}_i)\} f_u(t | \mathbf{X}_i) \quad (1.41)$$

Consequentemente, a função de risco populacional pode ser escrita como:

$$h(t | \mathbf{X}_i, \mathbf{Z}_i) = \frac{\{1 - \rho(\mathbf{Z}_i)\} f_u(t | \mathbf{X}_i)}{\rho(\mathbf{Z}_i) + \{1 - \rho(\mathbf{Z}_i)\} S_u(t | \mathbf{X}_i)} \quad (1.42)$$

A principal vantagem dos modelos de mistura é a interpretação direta dos componentes. O modelo de incidência permite identificar características associadas à probabilidade de cura, enquanto o modelo de latência permite avaliar os fatores relacionados ao tempo até o evento entre os indivíduos que permanecem suscetíveis. Essa separação deve ser considerada cuidadosamente na interpretação dos coeficientes. Um efeito positivo no componente de incidência pode representar maior chance de cura, enquanto um efeito positivo no componente de latência pode representar maior risco ou menor sobrevivência entre os suscetíveis, dependendo da parametrização adotada. Portanto, coeficientes pertencentes aos dois componentes não devem ser interpretados da mesma maneira.

A identificação da fração de cura também depende da duração e da qualidade do acompanhamento. Quando o último evento é seguido por um período muito curto ou quando poucos indivíduos permanecem sob risco nos tempos finais, pode ser difícil distinguir uma subpopulação verdadeiramente curada de indivíduos suscetíveis com tempos de sobrevivência muito longos (Maller and Zhou, 1996; Othus et al., 2012). Além disso, uma elevada proporção de censura pode produzir grande incerteza sobre a classificação latente

dos indivíduos. O fato de uma observação ser censurada não significa que o indivíduo seja curado. A estimação do estado latente deve decorrer da combinação entre o tempo de acompanhamento, o padrão dos eventos, as covariáveis e a estrutura probabilística especificada pelo modelo.

Métodos semiparamétricos para regressão com fração de cura foram desenvolvidos com o objetivo de reduzir a dependência de distribuições paramétricas rígidas para o componente de latência (Peng, 2003). Também foram propostas estruturas paramétricas flexíveis para aplicações em estudos populacionais de câncer, permitindo representar formas mais complexas das funções de risco e sobrevivência (Andersson et al., 2011). Outras extensões incorporam efeitos não lineares das covariáveis na presença de uma fração de cura (Ramires et al., 2018). A modelagem de cura também foi ampliada para contextos nos quais medidas longitudinais e tempos de sobrevivência são analisados conjuntamente, permitindo relacionar a evolução de biomarcadores ao componente de sobrevivência com fração de cura (Yang et al., 2021).

1.3.2 Modelos de Cura de Não-Mistura

Nos modelos de cura de não-mistura, a função de sobrevivência populacional não é apresentada diretamente como uma combinação convexa entre as funções de sobrevivência das subpopulações curada e suscetível. Em vez disso, a fração de cura surge como consequência da especificação de uma função de sobrevivência imprópria ou de um mecanismo latente responsável pela ocorrência do evento. Uma das principais construções dessa classe é o modelo de tempo de promoção, fundamentado na existência de um número latente de causas capazes de produzir o evento (Tsodikov et al., 2003). Em aplicações oncológicas, essas causas podem ser interpretadas, por exemplo, como células ou focos tumorais residuais que permanecem após o tratamento e podem desencadear uma recorrência.

Para uma melhor compreensão deste modelo, considere N_i como o número latente de causas iniciadoras do evento para o indivíduo i . Condicionamente a $N_i = n$, sejam $W_{i1}, \dots, W_{in}$ os tempos necessários para que cada causa produza o evento. O tempo observado até o evento, neste caso, é definido como:

$$T_i = \min(W_{i1}, \dots, W_{iN_i}) \quad (1.43)$$

Adota-se a convenção de que $T_i = \infty$ quando $N_i = 0$, situação na qual o indivíduo é considerado curado ou não suscetível ao evento. Agora, suponha que, condicionalmente às covariáveis, N_i siga uma distribuição de Poisson com média $\kappa(\mathbf{Z}_i)$, isto é, $N_i | \mathbf{Z}_i \sim \text{Poisson}\{\kappa(\mathbf{Z}_i)\}$ e que os tempos latentes W_{ij} sejam independentes e identicamente distribuídos, com função de sobrevivência $S_0(t | \mathbf{X}_i)$. Condicionalmente a $N_i = n$, a função de sobrevivência do menor dos tempos latentes é dada por:

$$P(T_i > t | N_i = n, \mathbf{X}_i) = \{S_0(t | \mathbf{X}_i)\}^n \quad (1.44)$$

Ao calcular a esperança em relação à distribuição de N_i , obtém-se a função de sobrevivência populacional,

$$S(t | \mathbf{X}_i, \mathbf{Z}_i) = \exp[-\kappa(\mathbf{Z}_i) \{1 - S_0(t | \mathbf{X}_i)\}] \quad (1.45)$$

de modo que se $S_0(t \mid \mathbf{X}_i) \rightarrow 0$ quando $t \rightarrow \infty$, então $\lim_{t \rightarrow \infty} S(t \mid \mathbf{X}_i, \mathbf{Z}_i) = \exp \{-\kappa(\mathbf{Z}_i)\}$. Assim, a fração de cura é dada por:

$$\rho(\mathbf{Z}_i) = P(N_i = 0 \mid \mathbf{Z}_i) = \exp \{-\kappa(\mathbf{Z}_i)\} \quad (1.46)$$

Na presença de covariáveis, uma especificação usual relaciona o número médio de causas às covariáveis por $\kappa(\mathbf{Z}_i) = \exp(\mathbf{Z}_i^\top \boldsymbol{\zeta})$. Nesse caso, a probabilidade de cura é:

$$\rho(\mathbf{Z}_i) = \exp \{-\exp(\mathbf{Z}_i^\top \boldsymbol{\zeta})\} \quad (1.47)$$

e a ligação correspondente pode ser expressa como:

$$\log [-\log \{\rho(\mathbf{Z}_i)\}] = \mathbf{Z}_i^\top \boldsymbol{\zeta} \quad (1.48)$$

A interpretação dos coeficientes difere daquela adotada no modelo de mistura com ligação logística. No modelo de tempo de promoção, um valor positivo de ζ_j aumenta o número esperado de causas iniciadoras e, conseqüentemente, reduz a probabilidade de cura. Por essa razão, a direção dos efeitos deve ser interpretada de acordo com a parametrização específica do componente de incidência. A formulação por tempo de promoção fornece uma interpretação mecanicista para a ocorrência do evento e permite que a cura surja naturalmente quando $N_i = 0$. Extensões podem substituir a distribuição de Poisson por outras distribuições de contagem, permitindo diferentes níveis de dispersão e produzindo classes mais amplas de modelos de sobrevivência de longo prazo (Tsodikov et al., 2003).

Embora os modelos de mistura e não-mistura apresentem parametrizações distintas, ambas as classes produzem funções de sobrevivência populacionais com limite positivo. A escolha entre elas deve considerar o objetivo da análise, a interpretação científica desejada, a adequação aos dados e a estabilidade inferencial. Nesta dissertação, a estrutura de mistura assume papel central, pois permite modelar separadamente o componente de incidência, associado à probabilidade de cura, e o componente de latência, associado à distribuição do tempo até o evento entre os suscetíveis. O componente de latência é descrito por meio de um modelo semiparamétrico de transformação do risco acumulado, possibilitando acomodar diferentes relações entre as covariáveis e a sobrevivência.

1.4 Organização da Dissertação

Além deste capítulo introdutório, esta dissertação está organizada em mais cinco capítulos, estruturados de modo a apresentar, sequencialmente, os objetivos do trabalho, a formulação metodológica proposta, sua avaliação por meio de simulação, as aplicações em dados reais e as considerações finais.

O Capítulo 2, por exemplo, apresenta o objetivo geral e os objetivos específicos que orientam o desenvolvimento da pesquisa. Os objetivos delimitam os aspectos metodológicos, inferenciais, computacionais e aplicados considerados na construção e na avaliação da abordagem proposta.

Já o Capítulo 3 apresenta a formulação do modelo Bayesiano semiparamétrico de transformação do risco acumulado com fração de cura e o desenvolvimento do correspondente procedimento de inferência Bayesiana. Inicialmente, são definidos os componentes de incidência e latência, responsáveis, respectivamente, pela modelagem da probabilidade de cura e da distribuição do tempo até o evento entre os indivíduos suscetíveis. Em seguida, são

apresentadas as funções de transformação previamente especificadas e a transformação monótona adaptativa aos dados, bem como a representação não paramétrica e penalizada do risco acumulado *baseline*, a função de verossimilhança e as distribuições a priori atribuídas aos parâmetros e às funções desconhecidas. O capítulo também contempla a construção da distribuição a posteriori, a descrição do algoritmo Metropolis-within-Gibbs utilizado para a obtenção das amostras a posteriori e a definição do DIC e do WAIC para a comparação das diferentes especificações do modelo.

O Capítulo 4, por sua vez, descreve o estudo de simulação de Monte Carlo conduzido para avaliar o desempenho do modelo com transformação monótona adaptativa em amostras finitas. São apresentados o mecanismo de geração dos dados, os tamanhos amostrais, os valores verdadeiros dos parâmetros, a configuração da função de transformação, o risco acumulado *baseline* e o mecanismo adotado para a geração dos tempos de censura. O desempenho inferencial é avaliado por meio do viés, da raiz do erro quadrático médio, da cobertura empírica e do comprimento dos intervalos de credibilidade. Também são avaliadas a recuperação da função de transformação e do risco acumulado *baseline*, a dimensão resultante do modelo e as características computacionais do procedimento de estimação, incluindo taxas de aceitação, tamanho efetivo da amostra e tempo de execução.

O Capítulo 5 apresenta a aplicação da abordagem aos dados de câncer de mama provenientes de uma coorte retrospectiva do Netherlands Cancer Institute (NKI). São descritos o conjunto de dados, as características clínicas e tumorais consideradas e a especificação adotada para os componentes de incidência e latência. Em seguida, são apresentados os ajustes dos modelos Bayesianos com transformações de riscos proporcionais, chances proporcionais e monótona adaptativa, além do modelo Weibull com fração de cura e do modelo semiparamétrico de Cox com cura por mistura utilizado como referência frequentista. A análise contempla a interpretação dos efeitos das covariáveis nos componentes de incidência e latência, a estimação da função de transformação e do risco acumulado *baseline*, os diagnósticos das cadeias, a dependência a posteriori entre os componentes, a análise de sensibilidade às prioris e a comparação dos modelos por meio do DIC e do WAIC.

Por fim, o Capítulo 6 reúne as considerações finais da dissertação. Nesse capítulo, são sintetizados os principais resultados metodológicos, inferenciais e aplicados, discutidas as limitações da abordagem e apresentadas possíveis extensões para trabalhos futuros.

Objetivos

2.1 Geral

O presente trabalho tem por objetivo geral desenvolver e avaliar um *framework* Bayesiano semiparamétrico de transformação do risco acumulado para dados de sobrevivência com fração de cura, no qual os componentes de incidência e latência sejam estimados conjuntamente, o risco acumulado *baseline* seja representado de forma não paramétrica e penalizada e a relação entre as covariáveis e o risco possa ser definida por funções de transformação previamente especificadas ou por uma transformação monótona adaptativa estimada a partir dos dados.

2.2 Específicos

Para alcançar o objetivo geral, estabeleceu-se os seguintes objetivos específicos:

- formular um modelo Bayesiano semiparamétrico com fração de cura, distinguindo o componente de incidência, associado à probabilidade de cura ou sobrevivência de longo prazo, e o componente de latência, associado à distribuição do tempo até o evento entre os indivíduos suscetíveis;
- incorporar ao componente de latência funções de transformação do risco acumulado previamente especificadas, incluindo estruturas de riscos proporcionais e de chances proporcionais;
- desenvolver uma função de transformação monótona adaptativa aos dados, construída por meio de funções de base *spline* e estimada em conjunto com os demais componentes do modelo, permitindo representar afastamentos das estruturas de transformação previamente especificadas;
- representar o risco acumulado *baseline* de forma não paramétrica por meio de incrementos positivos definidos nos tempos de evento observados, com regularização baseada em uma penalização de passeio aleatório de segunda ordem na escala logarítmica;
- incorporar uma penalização de passeio aleatório de segunda ordem aos coeficientes da transformação adaptativa, controlando a suavidade da função de transformação estimada;

- desenvolver um procedimento de inferência Bayesiana baseado em um algoritmo Metropolis-within-Gibbs para a estimação conjunta dos coeficientes de incidência e latência, dos coeficientes da transformação adaptativa, dos incrementos do risco acumulado *baseline* e dos parâmetros de suavização;
- avaliar, por meio de um estudo de simulação de Monte Carlo, o desempenho do modelo adaptativo em diferentes tamanhos amostrais e níveis de censura, considerando viés, raiz do erro quadrático médio, cobertura empírica e comprimento dos intervalos de credibilidade, além da recuperação das funções de transformação e de risco acumulado *baseline* e do comportamento computacional do procedimento de estimação;
- aplicar os modelos propostos a dados de câncer de mama do Netherlands Cancer Institute, avaliando os efeitos das covariáveis nos componentes de incidência e latência, a função de transformação estimada, o risco acumulado *baseline*, a sensibilidade às prioris e a dependência a posteriori entre os componentes do modelo;
- comparar as especificações Bayesiana adaptativa, de riscos proporcionais, de chances proporcionais e Weibull com fração de cura por meio dos critérios DIC e WAIC, utilizando também o modelo semiparamétrico de Cox com cura por mistura como referência frequentista.

Framework Bayesiano Para Modelos Semiparamétrico de Transformação do Risco com Fração de Cura

Neste capítulo é apresentado um *framework* que integra três componentes de modelagem tratados de forma separada na literatura de análise de sobrevivência: um modelo flexível de transformação do risco acumulado, uma estrutura de cura por mistura e uma especificação semiparamétrica Bayesiana para o risco *baseline*. Em contraste com os modelos clássicos de cura com riscos proporcionais, o parâmetro de transformação θ é estimado a partir dos dados, permitindo a seleção adaptativa da relação estrutural de risco. Ao mesmo tempo, o risco acumulado *baseline* é modelado de forma não paramétrica por meio de incrementos positivos suaves, proporcionando maior flexibilidade do que as alternativas paramétricas ou exponenciais por partes convencionais. Essa formulação unificada permite realizar inferência sobre incidência, latência, transformação estrutural e risco *baseline* em um único *framework* computacional.

3.1 Especificação do Modelo de Transformação

Seja T um tempo de falha não negativo e $\mathbf{X} \in \mathbb{R}^l$ o vetor de planejamento associado à latência. Considera-se um *framework* de transformação do risco acumulado (Zeng and Lin, 2006), no qual a função de risco acumulado condicional é modelada como,

$$\Lambda(t | \mathbf{X}) = G\left(e^{\beta^\top \mathbf{X}} \Lambda_0(t); \theta\right), \quad (3.1)$$

em que $\Lambda_0(t)$ é uma função de risco acumulado *baseline* não especificada, $\beta \in \mathbb{R}^l$ é um vetor de parâmetros de regressão e $G(\cdot; \theta)$ é uma função de transformação conhecida, crescente em sentido estrito e indexada por um parâmetro θ . A função de transformação $G(\cdot; \theta)$ determina a relação estrutural entre as covariáveis e o risco. Essa formulação inclui, como casos particulares, diversos modelos utilizados em análise de sobrevivência, entre eles:

- Riscos proporcionais: $G(x) = x$.
- Chances proporcionais: $G(x) = \log(1 + x)$.
- Modelos de transformação Box-Cox: $G(x; \theta) = [(1 + x)^\theta - 1]/\theta$, $\theta \neq 0$.

- Modelos de transformação exponencial: $G(x) = e^x$.
- Modelos de transformação logarítmica: $G(x; \theta) = (1/\theta) \log(1 + \theta x)$, $\theta \geq 0, \theta \neq 1$.

Diferentemente das formulações semiparamétricas clássicas de riscos proporcionais, o *framework* de transformação permite desvios flexíveis em relação às estruturas de risco multiplicativas, preservando a interpretabilidade por meio de uma transformação monótona do risco (Chen and Lu, 2012; Demarqui et al., 2019). Essa flexibilidade é relevante em contextos de sobrevivência de longo prazo, nos quais pressuposições estruturais fixas sobre as razões de risco podem ser restritivas. Neste trabalho, o parâmetro de transformação θ é tratado como desconhecido e estimado em conjunto com (β, Λ_0) , permitindo que os dados determinem a forma estrutural apropriada dentro da classe de transformação.

3.2 Extensão para Sobrevivência de Longo Prazo

Para acomodar sobreviventes de longo prazo, preservando uma separação coerente entre incidência e latência, utilizamos $\mathbf{X}_i$ para o vetor de planejamento da latência e $\mathbf{Z}_i$ para o vetor de planejamento da incidência. Os preditores de interesse podem se sobrepor, sendo que o vetor de planejamento da latência é especificado sem intercepto, enquanto o vetor de planejamento da incidência inclui seu intercepto. Essa parametrização ancora a escala da latência ao risco acumulado *baseline*. Adotamos uma representação unificada de causas competitivas (Rodrigues et al., 2009). Para o indivíduo i com covariáveis $\mathbf{Z}_i$, seja N_i o número latente de potenciais causas iniciadoras, com probabilidades condicionais $p_{in} = P(N_i = n \mid \mathbf{Z}_i)$ para $n = 0, 1, 2, \dots$. Condicionalmente a $N_i = n \geq 1$, sejam $\{W_{i1}, \dots, W_{in}\}$ tempos latentes específicos de cada causa, i.i.d. independentes de N_i dado $\mathbf{Z}_i$, compartilhando uma função de sobrevivência condicional comum $S_0(t \mid \mathbf{X}_i)$ que governa o mecanismo de latência. O tempo observado até o evento é definido como $T_i = \min(W_{i1}, \dots, W_{iN_i})$, com a convenção $T_i = \infty$ quando $N_i = 0$. Sob essa construção, a função de sobrevivência populacional satisfaz

$$S(t \mid \mathbf{X}_i, \mathbf{Z}_i) = P(T_i > t \mid \mathbf{X}_i, \mathbf{Z}_i) = \sum_{n=0}^{\infty} p_{in} \{S_0(t \mid \mathbf{X}_i)\}^n = A_i(S_0(t \mid \mathbf{X}_i)), \quad (3.2)$$

em que $A_i(s) = \sum_{n=0}^{\infty} p_{in} s^n$ é a função geradora de probabilidades (fgp) de N_i . Se $S_0(t \mid \mathbf{X}_i) \rightarrow 0$ quando $t \rightarrow \infty$, então

$$\lim_{t \rightarrow \infty} S(t \mid \mathbf{Z}_i) = A_i(0) = p_{i0} = P(N_i = 0 \mid \mathbf{Z}_i), \quad (3.3)$$

de modo que p_{i0} representa a fração de longo prazo (cura/imunidade). Essa formulação resulta em uma classe ampla de modelos de sobrevivência de longo prazo mediante a especificação de (i) um modelo de latência $S_0(\cdot)$ e (ii) um modelo de incidência por meio da fgp $A_i(\cdot)$. Em nosso *framework*, o mecanismo de latência é especificado por meio do modelo de transformação do risco acumulado introduzido na Seção 3.1. Em particular,

$$S_0(t \mid \mathbf{X}_i) = \exp \left\{ -G \left(e^{\beta^\top \mathbf{X}_i} \Lambda_0(t); \theta \right) \right\}, \quad (3.4)$$

em que $\Lambda_0(t)$ é um risco acumulado *baseline* não especificado e $G(\cdot; \theta)$ é crescente em sentido estrito. Como consequência, uma vez definido $S_0(t \mid \mathbf{X}_i)$, a sobrevivência populacional é obtida pelo mapeamento $S(t \mid \mathbf{X}_i, \mathbf{Z}_i) = A_i\{S_0(t \mid \mathbf{X}_i)\}$, e diferentes modelos de

sobrevivência de longo prazo correspondem a diferentes escolhas paramétricas para $A_i(\cdot)$ e para sua dependência das covariáveis. Dois submodelos importantes decorrem de (3.2). Primeiro, o *modelo clássico de cura por mistura* é obtido ao restringir $N_i \in \{0, 1\}$, de modo que $p_{i0} = \rho(\mathbf{Z}_i)$ e $p_{i1} = 1 - \rho(\mathbf{Z}_i)$. Nesse caso, $A_i(s) = \rho(\mathbf{Z}_i) + \{1 - \rho(\mathbf{Z}_i)\}s$ e

$$S(t \mid \mathbf{X}_i, \mathbf{Z}_i) = \rho(\mathbf{Z}_i) + \{1 - \rho(\mathbf{Z}_i)\}S_0(t \mid \mathbf{X}_i), \quad (3.5)$$

em que $\rho(\mathbf{Z}_i)$ é interpretado como a probabilidade de cura. Relacionamos $\rho(\mathbf{Z}_i)$ às covariáveis por meio da regressão logística

$$\rho(\mathbf{Z}_i) = \frac{\exp(\boldsymbol{\gamma}^\top \mathbf{Z}_i)}{1 + \exp(\boldsymbol{\gamma}^\top \mathbf{Z}_i)}, \quad (3.6)$$

com $\boldsymbol{\gamma}$ governando o componente de incidência. Segundo, uma alternativa flexível que inclui comportamentos de longo prazo do tipo promoção e geométrico é obtida ao assumir um modelo de incidência binomial negativo para N_i , com uma fgp da forma

$$A_i(s) = \{1 + \alpha \kappa_i(1 - s)\}^{-1/\alpha}, \quad \kappa_i = \exp(\boldsymbol{\gamma}^\top \mathbf{Z}_i), \quad \alpha > 0, \quad (3.7)$$

que resulta em

$$S(t \mid \mathbf{X}_i, \mathbf{Z}_i) = \{1 + \alpha \kappa_i(1 - S_0(t \mid \mathbf{X}_i))\}^{-1/\alpha}, \quad p_{i0} = \{1 + \alpha \kappa_i\}^{-1/\alpha}. \quad (3.8)$$

O modelo de tempo de promoção surge como um caso limite quando $\alpha \rightarrow 0$, resultando em $A_i(s) = \exp\{-\kappa_i(1 - s)\}$ e $S(t \mid \mathbf{X}_i, \mathbf{Z}_i) = \exp\{-\kappa_i(1 - S_0(t \mid \mathbf{X}_i))\}$ com fração de cura $p_{i0} = \exp(-\kappa_i)$. O modelo geométrico de sobrevivência de longo prazo é obtido como um caso particular de (3.7) (por exemplo, $\alpha = 1$ sob essa parametrização), resultando em $S(t \mid \mathbf{X}_i, \mathbf{Z}_i) = \{1 + \kappa_i(1 - S_0(t \mid \mathbf{X}_i))\}^{-1}$ e fração de cura $p_{i0} = (1 + \kappa_i)^{-1}$. Neste trabalho, adotamos a especificação do modelo de cura por mistura (3.5)–(3.6), pois ela preserva a decomposição usual e de interpretação clínica entre incidência (probabilidade de cura) e latência (sobrevivência dos suscetíveis), ao mesmo tempo em que permite que o componente de latência (3.4) se afaste da suposição de riscos proporcionais por meio da transformação $G(\cdot; \theta)$ e de um risco acumulado *baseline* não paramétrico $\Lambda_0(t)$. No entanto, a formulação geral (3.2) mostra que escolhas alternativas para $A_i(\cdot)$ (por exemplo, tempo de promoção, binomial negativa e geométrica) podem ser adotadas quando forem mais adequadas do ponto de vista científico, com o desenvolvimento posterior da especificação semiparamétrica Bayesiana penalizada para o risco *baseline* e da inferência posterior seguindo de forma análoga após a substituição do mapeamento de $S_0(t \mid \mathbf{X}_i)$ para $S(t \mid \mathbf{X}_i, \mathbf{Z}_i)$.

3.3 Especificação da Priori Penalizada

Nas Seções 3.1–3.2, o modelo de transformação com fração de cura foi especificado por meio de $\boldsymbol{\beta}$ (componente de latência), $\boldsymbol{\gamma}$ (componente de incidência), do parâmetro de transformação estrutural θ e de um risco acumulado *baseline* não especificado $\Lambda_0(t)$. Completamos agora a especificação do modelo por meio da formulação de uma priori Bayesiana semiparamétrica para $\Lambda_0(t)$ e da obtenção da distribuição a posteriori correspondente. Em particular, modelamos $\Lambda_0(t)$ como uma função escada contínua à direita,

com saltos nos tempos de evento distintos e ordenados $t_{(1)} < \dots < t_{(J)}$:

$$\Lambda_0(t) = \sum_{j=1}^J \gamma_j I(t \geq t_{(j)}), \quad \gamma_j > 0. \quad (3.9)$$

O incremento $\gamma_j = \Delta\Lambda_0(t_{(j)})$ representa o salto do risco acumulado *baseline* em $t_{(j)}$. Essa formulação tem caráter não paramétrico e é semelhante ao estimador de máxima verossimilhança não paramétrico dos coeficientes de regressão do modelo de Cox (Wang, 2008), ao mesmo tempo em que permite regularização Bayesiana. Observe que, para $t \in (t_{(j-1)}, t_{(j)}]$, o risco acumulado *baseline* satisfaz,

$$\Lambda_0(t_{(j)}) = \sum_{k=1}^j \gamma_k, \quad \text{e} \quad \Lambda_0(t_{(j-1)}) = \sum_{k=1}^{j-1} \gamma_k. \quad (3.10)$$

De forma equivalente, a medida ajustada do risco acumulado *baseline* é atômica nos tempos de evento observados,

$$d\Lambda_0(t) = \sum_{j=1}^J \gamma_j \delta_{t_{(j)}}(dt), \quad (3.11)$$

em que $\delta_{t_{(j)}}$ é uma massa pontual unitária. Esse suporte fornece a representação não paramétrica usual da verossimilhança nos tempos de evento observados. A simulação gera tempos de falha contínuos, enquanto a estimação adapta seu suporte às falhas distintas observadas em cada conjunto de dados. Considere a função de sobrevivência dos suscetíveis $S_0(t \mid \mathbf{X})$ apresentada na Seção 3.2. Para o indivíduo i , defina $\eta_i = \boldsymbol{\beta}^\top \mathbf{X}_i$ e

$$U_{ij} = e^{\eta_i} \Lambda_0(t_{(j)}) \quad \text{e} \quad U_{i,j-1} = e^{\eta_i} \Lambda_0(t_{(j-1)}). \quad (3.12)$$

Como $\Lambda_0(t)$ é uma função escada, o risco acumulado transformado definido por $\Lambda(t \mid \mathbf{X}_i) = G(e^{\eta_i} \Lambda_0(t); \theta)$ também é uma função escada, com saltos em $t_{(j)}$. Portanto, o incremento do risco acumulado em $t_{(j)}$ para o indivíduo i é dado por

$$\Delta\Lambda(t_{(j)} \mid \mathbf{X}_i) = G(U_{ij}; \theta) - G(U_{i,j-1}; \theta). \quad (3.13)$$

Por construção, a função de sobrevivência dos suscetíveis satisfaz a relação $S_0(t \mid \mathbf{X}_i) = \exp\{-\Lambda(t \mid \mathbf{X}_i)\}$. Assim, a recursão entre dois tempos de evento consecutivos e ordenados é dada por,

$$S_0(t_{(j)} \mid \mathbf{X}_i) = S_0(t_{(j-1)} \mid \mathbf{X}_i) \exp\{-\Delta\Lambda(t_{(j)} \mid \mathbf{X}_i)\}. \quad (3.14)$$

Dessa forma, a probabilidade discreta de falha para um indivíduo suscetível no tempo $t_{(j)}$, condicionada à sobrevivência até $t_{(j-1)}$, é dada por,

$$P(T = t_{(j)} \mid T \geq t_{(j)}, \mathbf{X}_i) = 1 - \exp\{-\Delta\Lambda(t_{(j)} \mid \mathbf{X}_i)\}. \quad (3.15)$$

Segue que a massa de probabilidade dos suscetíveis em $t_{(j)}$ corresponde ao decréscimo da sobrevivência,

$$f_0(t_{(j)} \mid \mathbf{X}_i) = S_0(t_{(j-1)} \mid \mathbf{X}_i) [1 - \exp\{-\Delta\Lambda(t_{(j)} \mid \mathbf{X}_i)\}]. \quad (3.16)$$

Agora, sejam (T_i, δ_i) o tempo observado e o indicador de evento. Se $\delta_i = 1$ e $T_i = t_{(j)}$, a contribuição para a verossimilhança é a probabilidade de pertencer ao grupo suscetível multiplicada pela massa dos suscetíveis em $t_{(j)}$, isto é,

$$L_i = [1 - \rho(\mathbf{Z}_i)] S_0(t_{(j-1)} \mid \mathbf{X}_i) [1 - \exp\{-\Delta\Lambda(t_{(j)} \mid \mathbf{X}_i)\}]. \quad (3.17)$$

Por outro lado, se $\delta_i = 0$, a censura à direita implica que a contribuição para a verossimilhança é igual à sobrevivência populacional avaliada em T_i , isto é, $L_i = \rho(\mathbf{Z}_i) + [1 - \rho(\mathbf{Z}_i)] S_0(T_i \mid \mathbf{X}_i)$. Combinando os dois casos, a verossimilhança completa pode ser expressa como,

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \theta, \boldsymbol{\gamma}_0) = \prod_{i=1}^n \left\{ ([1 - \rho(\mathbf{Z}_i)] S_0(t_{(j(i)-1)} \mid \mathbf{X}_i) [1 - \exp\{-\Delta\Lambda(t_{(j(i))} \mid \mathbf{X}_i)\}])^{\delta_i} \times (\rho(\mathbf{Z}_i) + [1 - \rho(\mathbf{Z}_i)] S_0(T_i \mid \mathbf{X}_i))^{1-\delta_i} \right\}, \quad (3.18)$$

em que $j(i)$ denota o índice tal que $T_i = t_{(j(i))}$ quando $\delta_i = 1$. O vetor $\boldsymbol{\gamma}_0 = (\gamma_1, \dots, \gamma_J)$ representa os incrementos *baseline* e é mantido distinto dos coeficientes de incidência $\boldsymbol{\gamma}$ para evitar ambiguidade de notação. Agora, para garantir a positividade dos incrementos e preservar a flexibilidade não paramétrica, atribuímos prioris Gamma independentes, $\gamma_j \sim \text{Gamma}(a_0, b_0)$, $j = 1, \dots, J$, com média $E(\gamma_j) = a_0/b_0$ e variância $\text{Var}(\gamma_j) = a_0/b_0^2$. Essa escolha fornece uma aproximação discreta para uma priori de processo Gamma para $\Lambda_0(t)$. Além disso, para estabilizar a estimação e evitar irregularidade excessiva da trajetória *baseline* em amostras finitas, introduzimos uma penalização de suavidade sobre os logaritmos dos incrementos. Em particular, seja $\psi_j = \log(\gamma_j)$, e defina a diferença de segunda ordem,

$$\Delta^2 \psi_j = \psi_j - 2\psi_{j-1} + \psi_{j-2}, \quad j = 3, \dots, J. \quad (3.19)$$

Então, assumimos a penalização de passeio aleatório de segunda ordem (penalização RW2) (Rue and Held, 2005),

$$p(\boldsymbol{\psi} \mid \tau) \propto \exp \left\{ -\frac{\tau}{2} \sum_{j=3}^J (\Delta^2 \psi_j)^2 \right\}, \quad (3.20)$$

que contrai $\{\log \gamma_j\}$ em direção a uma tendência linear local e, com isso, impõe variação suave de $\{\gamma_j\}$ na escala multiplicativa. A combinação da priori Gamma com (3.20) resulta no núcleo da priori penalizada para o componente *baseline*

$$p(\boldsymbol{\gamma}_0 \mid \tau) \propto \left[\prod_{j=1}^J \gamma_j^{a_0-1} \exp(-b_0 \gamma_j) \right] \exp \left\{ -\frac{\tau}{2} \sum_{j=3}^J (\Delta^2 \log \gamma_j)^2 \right\}. \quad (3.21)$$

Uma hiperpriori Gamma é atribuída à precisão de suavização, isto é, $\tau \sim \text{Gamma}(a_\tau, b_\tau)$. Sob essa suposição, seja $\Theta = (\beta, \gamma, \theta, \gamma_0, \tau)$ e $\mathcal{D} = \{(T_i, \delta_i, \mathbf{X}_i, \mathbf{Z}_i)\}_{i=1}^n$. Então, a distribuição a posteriori conjunta é dada por,

$$p(\Theta | \mathcal{D}) \propto L(\beta, \gamma, \theta, \gamma_0) p(\beta) p(\gamma) p(\theta) p(\gamma_0 | \tau) p(\tau), \quad (3.22)$$

o que define um modelo Bayesiano no qual o risco acumulado *baseline* é modelado de forma não paramétrica por meio de incrementos positivos nos tempos de evento observados, enquanto a suavidade e a estabilidade numérica são controladas por uma priori penalizada que preserva a restrição de monotonicidade inerente aos riscos acumulados.

3.4 Inferência a Posteriori e Computação

A distribuição a posteriori obtida na Seção 3.3 não admite expressões em forma fechada para as distribuições marginais ou condicionais dos parâmetros. Em particular, os coeficientes de regressão β interagem de forma multiplicativa com o risco acumulado *baseline* $\Lambda_0(t)$, enquanto o parâmetro de transformação θ está inserido no mapeamento não linear $G(\cdot; \theta)$ e afeta cada incremento do risco. A penalização de suavidade introduz dependência estruturada entre incrementos *baseline* adjacentes. Assim, a inferência a posteriori é realizada por meio de um algoritmo de Monte Carlo via cadeias de Markov construído sob um *framework* Metropolis-within-Gibbs (MwG). Para implementar o algoritmo MwG, seja o vetor completo de parâmetros $\Theta = (\beta, \gamma, \theta, \gamma_0, \tau)$, em que $\gamma_0 = (\gamma_1, \dots, \gamma_J)$ denota os incrementos do risco acumulado *baseline* e τ é o parâmetro de precisão da suavização. Denotando a verossimilhança em (3.18) por $L(\Theta)$, a densidade log-posteriori em (3.22) é proporcional a,

$$\log p(\Theta | \mathcal{D}) \propto \log L(\Theta) + \log p(\beta) + \log p(\gamma) + \log p(\theta) + \log p(\gamma_0 | \tau) + \log p(\tau). \quad (3.23)$$

Como as distribuições condicionais de β , γ , θ e γ_0 não estão disponíveis em forma fechada, esses parâmetros são atualizados por meio de etapas do tipo Metropolis. Em particular,

- Os parâmetros de incidência γ são amostrados por meio de uma proposta Metropolis de passeio aleatório normal multivariada centrada no valor corrente, cuja escala escalar e matriz de covariância regularizada são estimadas a partir do histórico da cadeia durante o período de *burn-in* (Haario et al., 2001);
- os coeficientes de regressão da latência β também são atualizados por meio de uma etapa Metropolis multivariada no espaço de parâmetros sem restrições, utilizando o mesmo esquema de estimação da covariância durante o período de *burn-in*;
- os coeficientes livres de transformação θ são atualizados em conjunto por meio de uma etapa Metropolis multivariada, com a covariância da proposta estimada durante o período de *burn-in* e, em seguida, mantida fixa;
- Para os incrementos *baseline* γ_0 , a positividade e a estabilidade numérica são tratadas por meio da reparametrização na escala logarítmica, com $\psi_j = \log \gamma_j$. O vetor completo ψ é atualizado por meio de uma proposta de Langevin preconditionada

com ajuste de Metropolis (Roberts and Tweedie, 1996)

$$q(\psi' | \psi) = \mathcal{N}\left(\psi + \frac{\epsilon^2}{2} M \nabla_{\psi} \log p(\Theta | \mathcal{D}), \epsilon^2 M\right),$$

em que M é a inversa normalizada de $D_2^T D_2 + \kappa I_J$ para um valor fixo positivo de regularização κ . Esse termo de regularização controla a geometria da proposta e a eficiência da amostragem sem alterar a distribuição a posteriori alvo. O gradiente analítico inclui a verossimilhança do intervalo de evento, a contribuição da sobrevivência da mistura para observações censuradas, a priori dos incrementos e o Jacobiano, além da penalização RW2. A razão exata entre as propostas direta e reversa é incluída na probabilidade de aceitação de Metropolis, enquanto ϵ é adaptado durante o período de *burn-in* em direção à taxa de aceitação alvo do MALA de 0.574 (Roberts and Rosenthal, 1998).

- Por fim, para o parâmetro de precisão da suavização τ , considera-se uma etapa de Gibbs, uma vez que τ admite uma distribuição condicional completa conjugada. De fato, a combinação da hiperpriori Gamma com a forma quadrática induzida pela penalização de diferenças de segunda ordem resulta em,

$$\tau | - \sim \text{Gamma}\left(a_{\tau} + \frac{J-2}{2}, b_{\tau} + \frac{1}{2} \sum_{j=3}^J (\psi_j - 2\psi_{j-1} + \psi_{j-2})^2\right). \quad (3.24)$$

Assim, τ é atualizado por meio de uma etapa direta de Gibbs em cada iteração MCMC. Cada iteração MCMC consiste na atualização sequencial de γ , β , θ , ψ e τ .

Após o descarte de um período inicial de *burn-in*, os resumos a posteriori são obtidos a partir das amostras retidas. A convergência é avaliada por meio de gráficos de traço, diagnósticos de autocorrelação, $\hat{R}$ dividido e normalizado por postos, tamanho efetivo da amostra para a região central e tamanho efetivo da amostra para as caudas, considerando quatro cadeias inicializadas de forma independente (Vehtari et al., 2021). A inferência a posteriori para o risco acumulado *baseline* é obtida pela reconstrução de $\Lambda_0(t) = \sum_{t_{(j)} \leq t} e^{\psi_j}$, em cada iteração, resultando em intervalos de credibilidade para toda a trajetória *baseline*, bem como para as funções de sobrevivência e risco derivadas.

3.5 Especificação de uma Transformação Adaptativa aos Dados

A função de transformação $G(\cdot; \theta)$ introduzida na Seção 3.1 permite relações estruturais flexíveis entre as covariáveis e o risco acumulado. Escolhas clássicas, como riscos proporcionais, chances proporcionais ou transformações do tipo logarítmica, correspondem a formas paramétricas fixas. Embora convenientes, tais especificações podem ser restritivas quando a relação de risco subjacente se afasta dessas estruturas canônicas. Para ampliar a flexibilidade, preservando a monotonicidade e a interpretabilidade, introduzimos uma transformação adaptativa aos dados construída por meio de um gerador de splines monótonos. Em particular, definimos a função de transformação como,

$$G(x; \theta) = \int_0^x \exp\left\{\sum_{k=2}^K \theta_k \tilde{B}_k(\log(1+u))\right\} du, \quad x \geq 0, \quad (3.25)$$

em que $\{B_k(\cdot)\}_{k=1}^K$ denota um conjunto de funções de base spline. Utilizamos as funções centradas $\tilde{B}_k(v) = B_k(v) - B_k(0)$ e fixamos $\theta_1 = 0$ como coeficiente de referência, resultando nos $K-1$ coeficientes identificáveis $\boldsymbol{\theta}_{-1} = (\theta_2, \dots, \theta_K)$. Essa representação centrada de referência estabelece $G'(0) = 1$ e ancora a escala da transformação ao *baseline*. Por construção, a derivada de G é,

$$G'(x; \boldsymbol{\theta}) = \exp \left\{ \sum_{k=2}^K \theta_k \tilde{B}_k(\log(1+x)) \right\} > 0, \quad x \geq 0, \quad (3.26)$$

o que garante a monotonicidade estrita de $G(\cdot; \boldsymbol{\theta})$. Como consequência, o risco acumulado

$$\Lambda(t | \mathbf{X}) = G\left(e^{\boldsymbol{\beta}^\top \mathbf{X}} \Lambda_0(t); \boldsymbol{\theta}\right), \quad (3.27)$$

permanece não decrescente em t , assegurando a validade do modelo de sobrevivência. O uso de $\log(1+u)$ dentro da base spline estabiliza o comportamento próximo de zero e reduz a sensibilidade numérica para valores elevados de u . Além disso, modelos clássicos de transformação surgem como casos particulares de (3.25). Em particular, $\boldsymbol{\theta}_{-1} = \mathbf{0}$ resulta em $G(x) = x$, a transformação de riscos proporcionais, enquanto coeficientes estruturados não nulos podem aproximar a curvatura associada ao modelo de chances proporcionais. Assim, a formulação proposta estende os modelos clássicos, permitindo que os dados determinem a deformação estrutural apropriada do risco *baseline*. Para evitar sobreajuste e assegurar variação estrutural suave, impomos uma penalização de passeio aleatório de segunda ordem sobre os coeficientes spline. Seja $\boldsymbol{\theta}$ o vetor completo, com $\theta_1 = 0$ fixado, assumimos,

$$p(\boldsymbol{\theta} | \tau_\theta) \propto \exp \left\{ -\frac{\tau_\theta}{2} \sum_{k=3}^K (\theta_k - 2\theta_{k-1} + \theta_{k-2})^2 \right\}, \quad (3.28)$$

em que τ_θ é um parâmetro de precisão da suavização. Uma hiperpriori Gamma é atribuída a τ_θ , isto é, $\tau_\theta \sim \text{Gamma}(a_\theta, b_\theta)$. Sob essa especificação, a função de sobrevivência dos suscetíveis é definida por,

$$S_0(t | \mathbf{X}) = \exp \left\{ -G\left(e^{\boldsymbol{\beta}^\top \mathbf{X}} \Lambda_0(t); \boldsymbol{\theta}\right) \right\}, \quad (3.29)$$

e a formulação geral do modelo de cura da Seção 3.2 permanece inalterada. Todos os componentes são estimados em conjunto no *framework* Bayesiano descrito anteriormente, com $\boldsymbol{\theta}$ atualizado como um bloco adicional de parâmetros no esquema MCMC.

3.6 Critérios Bayesianos de Comparação de Modelos

Para comparar as diferentes especificações da função de transformação, são utilizados o critério de informação da *deviance* (DIC) (Spiegelhalter et al., 2002) e o critério de informação de Watanabe–Akaike (WAIC) (Watanabe, 2010). Esses critérios permitem comparar o ajuste dos modelos considerando a qualidade do ajuste e a complexidade associada à especificação adotada. Para construí-los, considere M amostras retidas da distribuição a posteriori após o descarte do período de *burn-in*. Seja $\boldsymbol{\Theta}^{(m)}$ o vetor de parâmetros obtido na iteração m , em que $\boldsymbol{\Theta}$ representa $\boldsymbol{\Theta}_F$ nos modelos com funções de

transformação previamente especificadas ou Θ_A no modelo com transformação monótona adaptativa. A função log-verossimilhança avaliada na m -ésima amostra a posteriori é denotada por

$$\ell^{(m)} = \ell \left(\Theta^{(m)} \mid \mathcal{D}_n \right). \quad (3.30)$$

A média a posteriori da função log-verossimilhança é definida por

$$\bar{\ell} = \frac{1}{M} \sum_{m=1}^M \ell^{(m)}. \quad (3.31)$$

Por sua vez, o vetor formado pelas médias a posteriori dos parâmetros é dado por

$$\bar{\Theta} = \frac{1}{M} \sum_{m=1}^M \Theta^{(m)}. \quad (3.32)$$

A função *deviance* associada ao modelo é definida por

$$D(\Theta) = -2\ell \left(\Theta \mid \mathcal{D}_n \right). \quad (3.33)$$

A média a posteriori da *deviance* é expressa por

$$\bar{D} = \frac{1}{M} \sum_{m=1}^M D \left(\Theta^{(m)} \right) = -2\bar{\ell}. \quad (3.34)$$

A *deviance* avaliada no vetor de médias a posteriori é dada por

$$D \left(\bar{\Theta} \right) = -2\ell \left(\bar{\Theta} \mid \mathcal{D}_n \right). \quad (3.35)$$

A complexidade efetiva do modelo é estimada pelo número efetivo de parâmetros

$$p_D = \bar{D} - D \left(\bar{\Theta} \right). \quad (3.36)$$

O critério de informação da *deviance* é definido por

$$\text{DIC} = \bar{D} + p_D. \quad (3.37)$$

O WAIC é definido a partir das contribuições individuais da verossimilhança. Seja

$$\ell_i^{(m)} = \log L_i \left(\Theta^{(m)} \right) \quad (3.38)$$

a contribuição do indivíduo i para a log-verossimilhança na m -ésima amostra a posteriori. A densidade preditiva pontual integrada é estimada por

$$\text{lppd} = \sum_{i=1}^n \log \left[\frac{1}{M} \sum_{m=1}^M \exp \left\{ \ell_i^{(m)} \right\} \right]. \quad (3.39)$$

A complexidade efetiva utilizada pelo WAIC é estimada pela soma das variâncias a posteriori das contribuições individuais da log-verossimilhança,

$$p_{\text{WAIC}} = \sum_{i=1}^n \text{Var}_{m=1, \dots, M} \left(\ell_i^{(m)} \right). \quad (3.40)$$

O critério de informação de Watanabe–Akaike é então calculado por

$$\text{WAIC} = -2 (\text{lppd} - p_{\text{WAIC}}). \quad (3.41)$$

No DIC, a penalização é determinada pela complexidade efetiva p_D , obtida a partir da distribuição a posteriori da *deviance*. No WAIC, a penalização é determinada por p_{WAIC} , calculado a partir da variabilidade a posteriori das contribuições individuais da log-verossimilhança. Para modelos ajustados ao mesmo conjunto de dados e avaliados sob a mesma representação da verossimilhança, menores valores de DIC e WAIC indicam melhor desempenho relativo. Os dois critérios são utilizados na comparação entre as funções de transformação previamente especificadas e, quando aplicável, a transformação monótona adaptativa. Seus valores são interpretados como medidas de comparação entre os modelos, e não como medidas absolutas da adequação do ajuste.

Estudo de Simulação

Neste capítulo, apresentamos um estudo de simulação de Monte Carlo conduzido para avaliar o desempenho em amostras finitas do modelo Bayesiano semiparamétrico penalizado de transformação do risco acumulado com fração de cura proposto. O estudo avalia: (i) vies dos parâmetros de regressão, (ii) raiz do erro quadrático médio (RMSE), (iii) cobertura empírica dos intervalos de credibilidade de 95% e (iv) estabilidade numérica e desempenho computacional.

4.1 Configuração do Estudo de Simulação

O mecanismo gerador dos dados seguiu as etapas abaixo:

1. **Covariáveis.** Para cada indivíduo $i = 1, \dots, n$, geramos $W_{i1} \sim \text{Bernoulli}(0.5)$ e $W_{i2} \sim \mathcal{N}(0, 1)$ de forma independente. O vetor de planejamento da latência foi $\mathbf{X}_i = (W_{i1}, W_{i2})^\top$, sem intercepto, enquanto o vetor de planejamento da incidência foi $\mathbf{Z}_i = (1, W_{i1}, W_{i2})^\top$. Essa parametrização ancora a escala da latência ao risco acumulado *baseline* não especificado.
2. **Incidência de Cura.** O indicador de cura foi amostrado de $C_i \sim \text{Bernoulli}(\rho_i)$, com probabilidade de incidência $\rho_i = \text{logit}^{-1}(Z_i^\top \gamma)$. Se $C_i = 1$, o indivíduo foi considerado curado e o tempo de falha latente foi definido como $T_i^{\text{fail}} = \infty$.
3. **Geração dos Tempos de Falha.** Para os indivíduos suscetíveis, os tempos de falha foram gerados de forma contínua a partir do modelo de transformação adaptativa proposto. Definimos $\Lambda_0(t) = \lambda_0 t$, com $\lambda_0 = 0.03$, geramos $E_i \sim \text{Exponential}(1)$ e utilizamos

$$T_i^{\text{fail}} = \frac{G_A^{-1}(E_i; \boldsymbol{\theta})}{\lambda_0 \exp(\mathbf{X}_i^\top \boldsymbol{\beta})}. \quad (4.1)$$

A inversa de G_A foi avaliada por interpolação monótona sobre o argumento do risco acumulado, complementada por busca exata de raiz na cauda extrema. Essa construção produz tempos de falha contínuos em todo o domínio temporal positivo.

4. **Mecanismo de Censura.** Tempos de censura aleatória independentes foram gerados a partir de uma distribuição exponencial cuja taxa foi calibrada numericamente, após integração sobre a distribuição das covariáveis dos suscetíveis, para

produzir 10% ou 30% de censura entre os indivíduos suscetíveis. Nenhuma censura administrativa foi imposta no estudo principal. Os dados observados foram $T_i = \min(T_i^{\text{fail}}, C_i^{\text{cens}})$ e $\delta_i = I(T_i^{\text{fail}} \leq C_i^{\text{cens}})$. Apresentamos a proporção de censura aleatória obtida entre os suscetíveis separada da proporção de cura e da proporção geral de censura observada.

5. **Distribuições a Priori.** Os coeficientes de latência receberam priors independentes $\mathcal{N}(0, 1)$. Ao intercepto de incidência foi atribuída uma priori $\mathcal{N}\{\text{logit}(0.15), 1\}$ e às duas inclinações de incidência, priors independentes $\mathcal{N}(0, 1)$. Os incrementos *baseline* receberam priors $\text{Gamma}(0.01, 0.01)$, e as precisões de suavização τ_ψ e τ_θ receberam hiperpriors $\text{Gamma}(1, 1)$. Priors de ancoragem independentes $\mathcal{N}(0, 1)$ e uma penalização RW2 foram atribuídas aos três coeficientes livres de transformação.
6. **Configuração da Simulação e Valores Verdadeiros dos Parâmetros.** O estudo utilizou $n \in \{30, 50, 100, 500\}$ e $R = 1000$ replicações para cada um dos dois níveis de censura. Os coeficientes verdadeiros foram $\beta = (0.35, -0.15)^\top$ e $\gamma = \{\text{logit}(0.15), -0.60, 0.25\}^\top$. A transformação adaptativa utilizou quatro funções B-spline cúbicas brutas e os três coeficientes identificáveis $\theta_{-1} = (0.55, 0.20, 0.15)^\top$. Em cada replicação ajustada, J foi determinado pelo número de falhas observadas distintas. O número de parâmetros resultante é $J + 10$, e sua razão em relação a n é apresentada para cada cenário.
7. **Computação a Posteriori.** A inferência a posteriori foi realizada utilizando um algoritmo Metropolis-within-Gibbs, com propostas multivariadas regularizadas de passeio aleatório, estimadas durante o período de *burn-in*, para (β, γ, θ) , atualizações MALA analíticas preconditionadas por RW2 para o vetor *log-baseline* ψ e etapas de amostragem de Gibbs para os parâmetros de precisão (τ_ψ, τ_θ) . A etapa MALA utiliza o gradiente completo da log-posteriori e a correção exata da proposta direta-reversa descrita na Seção 3.4. Cada cadeia final utilizou 2.000 iterações de *burn-in*/adaptação, seguidas por 5.000 amostras retidas. Além de viés, RMSE, cobertura e comprimento do intervalo, apresentamos a aceitação por bloco, o tamanho efetivo da amostra, o tempo de execução, a taxa de conclusão, o RMSE do risco acumulado *baseline*, o RMSE da função de transformação, a censura obtida, o valor ajustado de J e a razão entre o número de parâmetros e o tamanho amostral.

4.2 Critérios de Avaliação

Os resumos de Monte Carlo foram calculados ao longo das R replicações para avaliar as características operacionais do procedimento Bayesiano. Seja ϑ um parâmetro pertencente ao conjunto $\{\beta_j, \gamma_j, \theta_\ell\}$, e seja ϑ_0 seu valor verdadeiro. Para cada replicação $r = 1, \dots, R$, a média a posteriori $\hat{\vartheta}^{(r)}$ foi utilizada como estimador pontual. O viés foi calculado como

$$\text{Bias}(\vartheta) = \frac{1}{R} \sum_{r=1}^R \left(\hat{\vartheta}^{(r)} - \vartheta_0 \right), \quad (4.2)$$

enquanto a raiz do erro quadrático médio (RMSE) foi definida como

$$\text{RMSE}(\vartheta) = \left\{ \frac{1}{R} \sum_{r=1}^R \left(\hat{\vartheta}^{(r)} - \vartheta_0 \right)^2 \right\}^{1/2}. \quad (4.3)$$

A cobertura empírica dos intervalos de credibilidade a posteriori de 95% e o comprimento médio dos intervalos de credibilidade foram calculados como

$$\text{Coverage}(\vartheta) = \frac{1}{R} \sum_{r=1}^R \mathbf{1} \left(\vartheta_0 \in \left[\widehat{\vartheta}_{0.025}^{(r)}, \widehat{\vartheta}_{0.975}^{(r)} \right] \right), \quad (4.4)$$

e

$$\text{AvgLength}(\vartheta) = \frac{1}{R} \sum_{r=1}^R \left(\widehat{\vartheta}_{0.975}^{(r)} - \widehat{\vartheta}_{0.025}^{(r)} \right), \quad (4.5)$$

em que $\widehat{\vartheta}_{0.025}^{(r)}$ e $\widehat{\vartheta}_{0.975}^{(r)}$ denotam os quantis a posteriori de 2.5% e 97.5%, respectivamente. Além disso, o desempenho computacional foi avaliado por meio das taxas médias de aceitação de Metropolis ($\bar{\alpha}$) para cada bloco de parâmetros $(\beta, \gamma, \theta, \psi)$ e pelo tempo médio de execução. Como os coeficientes spline individuais dependem da base, a precisão da transformação também é avaliada por meio do RMSE integrado em grade da função estimada $G(\cdot)$. A análise conjunta desse diagnóstico com o número de eventos observados, o valor ajustado de J , a precisão de suavização a posteriori, a aceitação e o ESS fornece uma caracterização da informação, da regularização e da eficiência computacional em cada cenário.

4.3 Resultados do Estudo de Simulação

4.3.1 Desempenho Computacional e Recuperação das Funções

Todas as 8.000 replicações foram concluídas, com 1.000 ajustes válidos em cada cenário e sem falha numérica terminal. O mecanismo de censura calibrado reproduziu os níveis estabelecidos: a censura aleatória média entre os indivíduos suscetíveis variou de 0.098 a 0.102 nos cenários com 10% de censura e de 0.299 a 0.301 nos cenários com 30% de censura. As proporções gerais correspondentes de censura observada foram de 0.207–0.211 e 0.383–0.387, respectivamente, enquanto as proporções de cura geradas variaram de 0.118 a 0.123. Como os tempos de falha foram contínuos, cada tempo de falha observado foi distinto e J foi igual ao número de eventos observados em cada replicação. A Tabela 4.1 apresenta o suporte resultante dos tempos de evento, a dimensão ajustada, a precisão no nível da função e o diagnóstico de ESS mais conservador.

O suporte adaptativo aos dados evita impor 60 incrementos *baseline* às amostras com $n = 30$ e $n = 50$. Embora a razão formal entre o número de parâmetros e o tamanho amostral tenha permanecido elevada, pois cada falha distinta gerou um salto *baseline*, a priori RW2 regularizou toda a trajetória e todos os ajustes permaneceram operacionais. As probabilidades médias de aceitação variaram de 0.288–0.362 para os blocos de regressão e transformação e de 0.612–0.715 para a atualização MALA do componente *baseline*. O ESS mínimo absoluto entre todas as replicações e diagnósticos registrados foi 53.7, e o menor ESS médio no nível dos cenários variou de 61.2 a 153.7. O tempo mediano de execução por replicação variou de 303 a 1.188 segundos. A precisão média de suavização a posteriori τ_ψ aumentou de 3.03–3.64 em $n = 30$ para 25.14–30.71 em $n = 500$, permitindo que a penalização se adaptasse ao aumento do suporte dos tempos de evento, enquanto τ_θ permaneceu entre 1.28 e 1.31.

Tabela 4.1: Configuração da simulação, dimensão ajustada e desempenho no nível da função. O ESS apresentado corresponde ao valor mínimo entre todas as replicações registradas e os blocos de diagnóstico.

Configuração da simulação e dimensão ajustada				
Nível de censura	n	Censura entre suscetíveis	Média de J (máx.)	$(J + 10)/n$
0.10	30	0.098	23.8 (29)	1.126
0.10	50	0.098	39.7 (48)	0.993
0.10	100	0.102	78.9 (93)	0.889
0.10	500	0.101	395.0 (426)	0.810
0.30	30	0.301	18.5 (27)	0.950
0.30	50	0.301	30.7 (42)	0.814
0.30	100	0.299	61.6 (78)	0.716
0.30	500	0.301	306.7 (340)	0.633
Desempenho no nível da função e diagnóstico computacional				
Nível de censura	n	RMSE($\hat{\Lambda}_0$)	RMSE($\hat{G}$)	ESS mínimo
0.10	30	1.142	1.713	66.6
0.10	50	1.145	1.942	70.0
0.10	100	1.121	2.196	60.4
0.10	500	0.769	2.064	53.7
0.30	30	1.038	1.530	62.9
0.30	50	0.952	1.714	63.9
0.30	100	0.969	1.989	62.9
0.30	500	0.822	2.305	55.1

4.3.2 Desempenho dos Parâmetros

Os coeficientes de latência foram recuperados com baixo viés absoluto em todos os cenários (no máximo 0.094), e seus RMSEs diminuíram entre $n = 30$ e $n = 500$. Para o componente de incidência, as duas inclinações associadas às covariáveis mantiveram cobertura entre 0.946 e 1.000; o intercepto, que controla de forma mais direta a separação na cauda, foi responsável pelo maior viés no nível do bloco e pela menor cobertura em $n = 500$. Os coeficientes brutos de transformação apresentaram contração persistente mesmo com a redução da amplitude de seus intervalos, indicando que a interpretação coeficiente a coeficiente é menos informativa quando uma transformação flexível e um componente *baseline* não paramétrico são estimados em conjunto. No nível da função, no entanto, o RMSE($\hat{G}$) permaneceu entre 1.530 e 2.305 na grade comum $[0, 10]$. Como a transformação verdadeira atinge $G_A(10) = 12.465$, isso corresponde a um RMSE normalizado da função entre 12.3% e 18.5%. Assim, a simulação sustenta a inferência por meio da função de transformação estimada e dos contrastes de sobrevivência específicos no tempo, enquanto os coeficientes brutos de spline atuam como coordenadas de base regularizadas.

Tabela 4.2: Características operacionais por bloco de parâmetros. São apresentados o viés absoluto máximo, o intervalo de RMSE e o intervalo de cobertura empírica de 95% entre os coeficientes pertencentes a cada bloco.

β				
Nível de censura	n	Viés absoluto máximo	RMSE	Cobertura 95%
0.10	30	0.045	0.292–0.467	0.956–0.961
0.10	50	0.059	0.227–0.381	0.942–0.960
0.10	100	0.085	0.159–0.282	0.932–0.959
0.10	500	0.094	0.079–0.159	0.892–0.903
0.30	30	0.054	0.345–0.487	0.945–0.974
0.30	50	0.051	0.262–0.401	0.940–0.960
0.30	100	0.048	0.172–0.305	0.955–0.957
0.30	500	0.079	0.088–0.169	0.928–0.931
γ				
Nível de censura	n	Viés absoluto máximo	RMSE	Cobertura 95%
0.10	30	0.626	0.378–0.702	0.986–0.998
0.10	50	0.693	0.388–0.757	0.956–0.998
0.10	100	0.710	0.376–0.761	0.926–0.992
0.10	500	0.765	0.269–0.834	0.595–0.960
0.30	30	0.477	0.372–0.578	0.996–1.000
0.30	50	0.542	0.369–0.621	0.992–1.000
0.30	100	0.640	0.368–0.706	0.972–0.995
0.30	500	0.740	0.294–0.799	0.827–0.978
θ				
Nível de censura	n	Viés absoluto máximo	RMSE	Cobertura 95%
0.10	30	0.728	0.390–0.732	0.999–1.000
0.10	50	0.771	0.450–0.778	0.990–1.000
0.10	100	0.815	0.558–0.833	0.869–1.000
0.10	500	0.749	0.473–0.821	0.406–0.991
0.30	30	0.712	0.326–0.715	1.000–1.000
0.30	50	0.745	0.377–0.750	1.000–1.000
0.30	100	0.801	0.451–0.813	0.961–1.000
0.30	500	0.832	0.562–0.899	0.422–0.985

Aplicação em Dados de Câncer de Mama

5.1 Introdução

Nas últimas décadas, os avanços no diagnóstico e no tratamento do câncer contribuíram para mudanças no prognóstico e na sobrevivência dos pacientes. Apesar desses avanços, a doença permanece entre as principais causas de mortalidade no mundo e constitui um problema de saúde pública que demanda métodos capazes de representar diferentes padrões de evolução clínica (Mariotto et al., 2014). De acordo com a International Agency for Research on Cancer (IARC), cerca de 20 milhões de novos casos de câncer e 9,7 milhões de mortes relacionadas à doença foram registrados em 2022, conforme dados publicados no Cancer Today, do Global Cancer Observatory (Ferlay et al., 2024). Entre os diferentes tipos de câncer, o câncer de mama (CM) é a neoplasia maligna mais diagnosticada entre mulheres, correspondendo a 11,6% dos novos casos globais e ocupando a quarta posição em mortalidade relacionada ao câncer, com 6,9% das mortes. No Brasil, o CM também apresenta relevância epidemiológica. Segundo estimativas do Instituto Nacional de Câncer (2025), foram estimados 73.610 novos casos de câncer de mama, além de 18.361 mortes associadas à doença, das quais 220 ocorreram em indivíduos do sexo masculino.

O câncer de mama compreende um conjunto heterogêneo de neoplasias originadas no tecido mamário, com diferenças quanto à origem histológica, ao perfil molecular, à evolução clínica e à resposta ao tratamento. Entre os principais tipos estão os carcinomas ductais, originados nos ductos mamários, e os carcinomas lobulares, originados nos lóbulos responsáveis pela produção de leite. Outros tipos menos frequentes incluem a doença de Paget da mama, o tumor filoide maligno e o angiossarcoma. Além da classificação histológica, os tumores mamários podem ser caracterizados segundo marcadores moleculares e expressão de receptores hormonais, características relacionadas ao prognóstico e à definição das estratégias terapêuticas. Em estágios iniciais, a doença pode ser assintomática, enquanto em fases mais avançadas podem ocorrer sinais como nódulos, alterações no tamanho ou formato da mama, modificações na pele, secreção mamilar e alterações na aréola (Sociedade Americana do Câncer (American Cancer Society), 2025).

A heterogeneidade da doença também se reflete nos fatores associados ao seu desenvolvimento e à evolução clínica. Entre os fatores de risco estão idade avançada, histórico familiar, mutações nos genes BRCA1 e BRCA2, fatores hormonais, obesidade, sedentarismo e outros aspectos relacionados ao estilo de vida (Foulkes, 2008). Características clínicas e tumorais, como tamanho do tumor, comprometimento linfonodal, grau histoló-

gico, presença de metástase e perfil molecular, também estão associadas ao prognóstico. O tratamento pode envolver cirurgia, radioterapia, quimioterapia, terapia hormonal, imunoterapia e terapias direcionadas, de acordo com o estágio da doença e as características biológicas do tumor. Entre os avanços terapêuticos, as terapias anti-HER2 contribuíram para alterações no prognóstico de subgrupos de pacientes (Cardoso et al., 2018).

As mudanças nos métodos de detecção precoce, como o rastreamento por mamografia, e nas estratégias terapêuticas contribuíram para o aumento da sobrevivência de pacientes com câncer de mama. Entretanto, os tempos de sobrevivência podem apresentar comportamentos distintos entre os indivíduos, em razão das características clínicas, tumorais e terapêuticas. Além disso, em estudos com acompanhamento prolongado, uma parcela dos pacientes pode permanecer livre do evento durante longos períodos, produzindo curvas de sobrevivência com estabilização em níveis superiores a zero. Esse comportamento é compatível com a presença de uma fração de indivíduos que não apresenta o evento de interesse durante o período de acompanhamento e motiva o uso de modelos de sobrevivência com fração de cura. Paralelamente, desigualdades no acesso ao diagnóstico e ao tratamento permanecem presentes, em especial em países de baixa e média renda, nos quais limitações no acesso aos serviços de saúde, ao diagnóstico precoce e ao tratamento adequado estão associadas a maiores taxas de mortalidade (Bray et al., 2018; Goldman and Schafer, 2011; McPherson et al., 2000; Kasper et al., 2016; Hunk et al., 2015).

Nesse contexto, a análise de dados de sobrevivência em câncer de mama requer modelos capazes de representar tanto a heterogeneidade entre pacientes quanto a presença de sobrevivência de longo prazo, sem impor uma única estrutura para a relação entre covariáveis e risco ao longo do tempo. Neste trabalho, os modelos Bayesianos semiparamétricos de transformação do risco acumulado com fração de cura são aplicados a dados de câncer de mama com o objetivo de avaliar os fatores associados à sobrevivência. A formulação considera que uma parcela dos indivíduos pode apresentar remissão completa ou permanecer livre do evento em longo prazo, caracterizando uma fração da população que pode não estar sujeita à ocorrência da falha. O modelo permite separar o componente de incidência, associado à probabilidade de cura ou sobrevivência de longo prazo, do componente de latência, associado à distribuição do tempo até o evento entre os indivíduos suscetíveis. Ao mesmo tempo, a transformação do risco acumulado permite flexibilizar a relação entre as covariáveis e o risco, enquanto a especificação semiparamétrica do risco *baseline* reduz a dependência de uma forma paramétrica previamente estabelecida.

5.2 Dados do Estudo

Para a aplicação da metodologia proposta, consideramos um conjunto de dados proveniente de uma coorte retrospectiva composta por 272 casos de carcinoma invasivo primário de mama, obtidos como dados secundários de van de Vijver et al. (2002), provenientes do Netherlands Cancer Institute (NKI). A coorte inclui pacientes do sexo feminino diagnosticadas entre 1984 e 1995, com os seguintes critérios de inclusão: diâmetro tumoral inferior a 5 cm no exame anatomopatológico (pT1 ou pT2), linfonodos axilares apicais negativos, idade ao diagnóstico de 52 anos ou menos e ausência de histórico prévio de câncer, exceto câncer de pele não melanoma. Todas as pacientes foram submetidas à mastectomia radical modificada ou à cirurgia conservadora da mama, com dissecação dos linfonodos axilares, seguida de radioterapia adjuvante quando indicada clinicamente. Apesar de o conjunto de dados ter origem no período de 1984–1995, ele mantém sua relevância devido à coleta abrangente e bem documentada de características clínicas e tumorais. Um resumo das características *baseline* e da distribuição do desfecho é apresentado na Tabela 5.1.

Tabela 5.1: Características *baseline* e dados operatórios da coorte retrospectiva analisada por [van de Vijver et al. \(2002\)](#).

Variáveis contínuas		
Variável	Não censuradas (77) Média $\pm$ DP	Censuradas (195) Média $\pm$ DP
Diâmetro	25.50 $\pm$ 8.64	21.35 $\pm$ 8.47
Linfonodos Positivos	1.65 $\pm$ 2.50	1.22 $\pm$ 1.93
Variáveis categóricas		
Variável	Não censuradas (77) n (%)	Censuradas (195) n (%)
Quimioterapia		
Não	51 (66%)	114 (58%)
Sim	26 (34%)	81 (42%)
Terapia Hormonal		
Não	71 (92%)	165 (85%)
Sim	6 (8%)	30 (15%)
Tipo de Cirurgia		
Conservadora	39 (51%)	113 (58%)
Mastectomia	38 (49%)	82 (42%)
Grau Tumoral		
I	4 (5%)	67 (34%)
II	24 (31%)	71 (37%)
III	49 (64%)	57 (29%)

Este conjunto de dados apresenta características que sustentam o uso do modelo Bayesiano semiparamétrico de transformação com prioris penalizadas proposto. A população do estudo é composta por pacientes com carcinoma invasivo primário de mama, apresentando variabilidade nas características tumorais e nos perfis de tratamento, conforme evidenciado pelas diferenças no diâmetro tumoral, número de linfonodos positivos, grau tumoral e estratégias terapêuticas, como quimioterapia, terapia hormonal e tipo de cirurgia, conforme apresentado na Tabela 5.1. A presença de variáveis contínuas (por exemplo, diâmetro tumoral e número de linfonodos positivos) e variáveis categóricas (por exemplo, modalidades de tratamento e características tumorais) sugere que os efeitos das covariáveis podem atuar por meio de mecanismos não lineares e não proporcionais. Em particular, as diferenças observadas entre os grupos censurado e não censurado indicam que a suposição de riscos proporcionais pode ser restritiva, motivando uma abordagem de modelagem mais flexível, na qual o componente de transformação permite a adaptação da função de risco acumulado a partir dos dados, relaxando a suposição de riscos proporcionais, enquanto a especificação semiparamétrica em degraus do risco *baseline* evita formas paramétricas restritivas. Além disso, o tamanho amostral moderado em relação ao número de preditores requer regularização para garantir estabilidade na estimação, obtida por meio da incorporação de prioris penalizadas, promovendo suavidade e contração.

5.3 Especificação e Ajuste dos Modelos

Na análise, então, para o componente semiparamétrico *baseline*, o suporte foi definido pelos tempos de evento distintos e ordenados $\{t_{(j)}\}_{j=1}^J$ obtidos a partir dos óbitos observados, com J igual ao número de tempos de evento únicos. Os incrementos *baseline* foram inicializados em $\gamma_{0j} = 0.02$ e receberam prioris Gamma fracamente informativas, com

hiperparâmetros $a_0 = b_0 = 0.01$. O componente de transformação foi representado por meio de um gerador de splines monótonos com $K = 4$ funções de base cúbicas avaliadas em $\log(1 + u)$, com o primeiro coeficiente bruto utilizado como referência zero e os três restantes estimados, e nós de fronteira fixados em $(0, \log\{1 + 2 \max(\text{OS})\})$ para estabilizar a extrapolação na escala transformada. O vetor de planejamento da latência utilizou o risco acumulado *baseline* como referência de escala e, portanto, incluiu os preditores clínicos sem intercepto. O diâmetro tumoral e o número de linfonodos positivos foram padronizados, de modo que seus coeficientes representam um aumento de um desvio padrão; os indicadores binários mantiveram sua codificação original 0/1. Aos coeficientes de regressão do bloco de latência foram atribuídas prioris independentes $\mathcal{N}(0, 1)$. Para o bloco de incidência, o intercepto foi centrado em uma probabilidade de cura *baseline* de ≈ 0.604 , correspondente ao nível empírico de sobrevivência de longo prazo sugerido pelo estimador de Kaplan–Meier (Figura 5.1), por meio de uma priori normal na escala logito, enquanto os demais coeficientes receberam prioris independentes $\mathcal{N}(0, 1)$. Prioris penalizadas foram impostas por meio de penalizações de passeio aleatório de segunda ordem sobre os coeficientes de transformação e sobre os logaritmos dos incrementos *baseline* $\psi_j = \log(\gamma_{0j})$, com as precisões de suavização τ_θ e τ_ψ recebendo hiperprioris Gamma $\text{Gamma}(1, 1)$. A sensibilidade às prioris foi avaliada por meio do reajuste do modelo adaptativo com $\tau_\theta, \tau_\psi \sim \text{Gamma}(0.1, 0.1)$ e da comparação dos resumos de regressão, do risco acumulado *baseline* e de $G(\cdot)$.

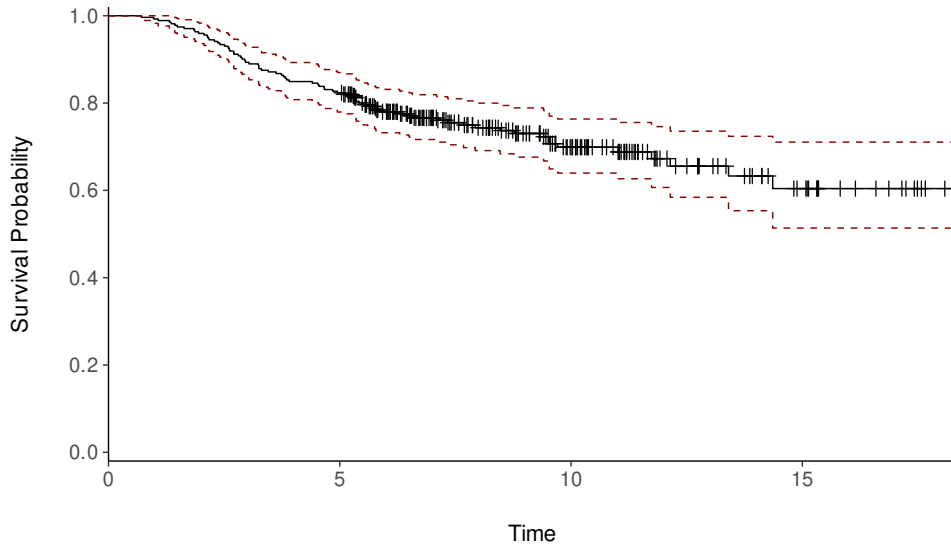


Figura 5.1: Curva de Kaplan–Meier para mulheres diagnosticadas com câncer de mama no conjunto de dados do NKI.

A inferência a posteriori foi conduzida utilizando um amostrador Metropolis-within-Gibbs. Os blocos de parâmetros (γ, β, θ) foram atualizados por meio de propostas Gaussianas de passeio aleatório, com matrizes de covariância multivariadas regularizadas estimadas durante o período de *burn-in*, e os incrementos *baseline* foram atualizados na escala logarítmica por meio da proposta MALA analítica preconditionada por RW2 descrita na Seção 3.4. As atualizações condicionais dos parâmetros de suavização (τ_ψ, τ_θ) foram realizadas por amostragem de Gibbs a partir de suas distribuições condicionais completas Gamma. Para cada especificação Bayesiana principal, foram executadas quatro cadeias inicializadas de forma independente com 40,000 iterações, das quais as primeiras 5,000 foram descartadas, e uma em cada cinco amostras após o período de *burn-in* foi retida,

resultando em 7,000 amostras retidas por cadeia. O ajuste adaptativo com a priori alternativa foi continuado a partir dos quatro estados armazenados, utilizando o mesmo esquema de 40,000 iterações, 5,000 iterações de *burn-in* e espaçamento entre amostras; os resumos apresentados utilizam as 7,000 amostras retidas da continuação de cada cadeia. A adaptação das propostas foi restrita ao período de *burn-in*; a escala do MALA teve como alvo uma probabilidade de aceitação de 0.574. A convergência é apresentada por meio do $\hat{R}$ dividido e normalizado por postos, ESS para a região central e ESS para as caudas, juntamente com gráficos de traço.

A aplicação compara o modelo adaptativo proposto com modelos de cura de riscos proporcionais e de chances proporcionais com funções de transformação fixas, um modelo Bayesiano Weibull de cura por mistura e o modelo semiparamétrico padrão de Cox com cura por mistura implementado no **smcure**. O WAIC e o DIC dos modelos Bayesianos utilizam a verossimilhança comum baseada nos intervalos de evento descrita após (3.18); o ajuste frequentista pelo **smcure** é apresentado em separado. Também apresentamos a matriz completa de correlações cruzadas a posteriori entre os coeficientes de latência e incidência. Para interpretar β quando G é estimada, apresentamos mudanças marginais médias na sobrevivência dos suscetíveis em 5 e 10 anos, utilizando perfis clínicos válidos para os contrastes entre fatores.

5.4 Resultados da Aplicação

5.4.1 Diagnósticos de Convergência

Os diagnósticos das quatro cadeias fornecem evidências de estabilidade na computação a posteriori (Tabela 5.2). Para o modelo adaptativo proposto, o valor máximo do $\hat{R}$ dividido e normalizado por postos foi 1.023 para os coeficientes de latência, 1.021 para os coeficientes de incidência e 1.005 para os coeficientes de transformação; os valores mínimos correspondentes de ESS para a região central foram 201, 196 e 1.140. O bloco *baseline* completo, com dimensão 77, também apresentou mistura adequada, com valor máximo de $\hat{R}$ igual a 1.049, ESS mínimo para a região central igual a 94 e ESS mínimo para as caudas igual a 232.

Tabela 5.2: Diagnósticos de convergência das quatro cadeias para os principais blocos de coeficientes na aplicação aos dados do NKI.

Modelos com Transformações Fixas				
Modelo	Bloco	ESS mínimo central	ESS mínimo das caudas	$\hat{R}$ máximo
PH	β	110	365	1.027
PH	γ	219	569	1.010
PO	β	124	264	1.022
PO	γ	98	718	1.040

Modelo Adaptativo e Modelo Weibull				
Modelo	Bloco	ESS mínimo central	ESS mínimo das caudas	$\hat{R}$ máximo
Adaptativo	β	201	346	1.023
Adaptativo	γ	196	600	1.021
Adaptativo	θ	1,140	2,449	1.005
Weibull	β	337	719	1.009
Weibull	γ	374	745	1.014

5.4.2 Inferência para o Modelo Adaptativo

Os resumos a posteriori para o modelo adaptativo são apresentados na Tabela 5.3. Entre as pacientes suscetíveis, o grau III foi associado à ocorrência mais precoce de eventos ($\beta = 1.556$, intervalo de credibilidade de 95%: 0.597 a 2.624). Como a transformação é estimada, em vez de fixada, esse coeficiente também é expresso na escala de sobrevivência: em relação ao grau de referência, a diferença média na sobrevivência das suscetíveis para o grau III foi de -0.314 em 5 anos (intervalo de credibilidade de 95%: -0.513 a -0.126) e de -0.413 em 10 anos (intervalo de credibilidade de 95%: -0.682 a -0.160). Esses contrastes específicos no tempo preservam a direção monótona de β , ao mesmo tempo em que fornecem uma medida de efeito com interpretação direta sob a transformação estimada.

Tabela 5.3: Resumos a posteriori para o modelo de cura com transformação adaptativa proposto na aplicação do NKI.

Bloco	Parâmetro	Média	DP	IC de 95%
Latência	Quimioterapia	-0.080	0.510	(-1.083, 0.866)
Latência	Hormonioterapia	-0.285	0.568	(-1.378, 0.846)
Latência	Mastectomia	0.028	0.386	(-0.765, 0.768)
Latência	Diâmetro tumoral (por DP)	0.356	0.227	(-0.118, 0.788)
Latência	Linfonodos positivos (por DP)	-0.007	0.207	(-0.378, 0.454)
Latência	Grau II	0.774	0.542	(-0.260, 1.903)
Latência	Grau III	1.556	0.505	(0.597, 2.624)
Incidência	Intercepto	0.167	0.774	(-1.548, 1.506)
Incidência	Quimioterapia	0.560	0.736	(-1.154, 1.776)
Incidência	Hormonioterapia	0.250	0.801	(-1.501, 1.667)
Incidência	Mastectomia	-0.162	0.578	(-1.451, 0.869)
Incidência	Diâmetro tumoral (por DP)	0.072	0.451	(-1.006, 0.968)
Incidência	Linfonodos positivos (por DP)	-0.468	0.391	(-1.217, 0.365)
Incidência	Grau II	-0.727	0.687	(-2.100, 0.611)
Incidência	Grau III	-1.043	0.651	(-2.324, 0.267)
Transformação	θ_2	-0.298	0.604	(-1.525, 0.862)
Transformação	θ_3	-0.269	0.680	(-1.596, 1.058)
Transformação	θ_4	-0.129	0.876	(-1.816, 1.598)
Suavização	τ_ψ	8.646	2.633	(4.404, 14.689)
Suavização	τ_θ	1.253	1.064	(0.101, 4.081)

A matriz completa de correlação a posteriori entre os coeficientes de latência e incidência é apresentada na Figura 5.2. A maior correlação, em valor absoluto, foi 0.648 para os coeficientes associados ao diâmetro tumoral, seguida por 0.626 para quimioterapia, 0.551 para mastectomia, 0.514 para grau II e 0.506 para linfonodos positivos. Esses resultados indicam dependência posterior entre os efeitos de algumas covariáveis nos componentes de incidência e latência, de modo que alterações no coeficiente associado ao tempo até o evento podem ser parcialmente compensadas por alterações no coeficiente associado à probabilidade de cura, refletindo a dificuldade dos dados em separar esses dois mecanismos com a mesma precisão. Esse comportamento é compatível com a limitada identificabilidade do componente de incidência, já sugerida pelos desvios-padrão a posteriori próximos aos das respectivas distribuições a priori. Ainda assim, todas as correlações permaneceram abaixo de 0.65, indicando que, embora exista dependência posterior moderada para algu-

mas covariáveis, não foram observadas associações próximas de uma dependência linear completa entre os coeficientes de incidência e latência.

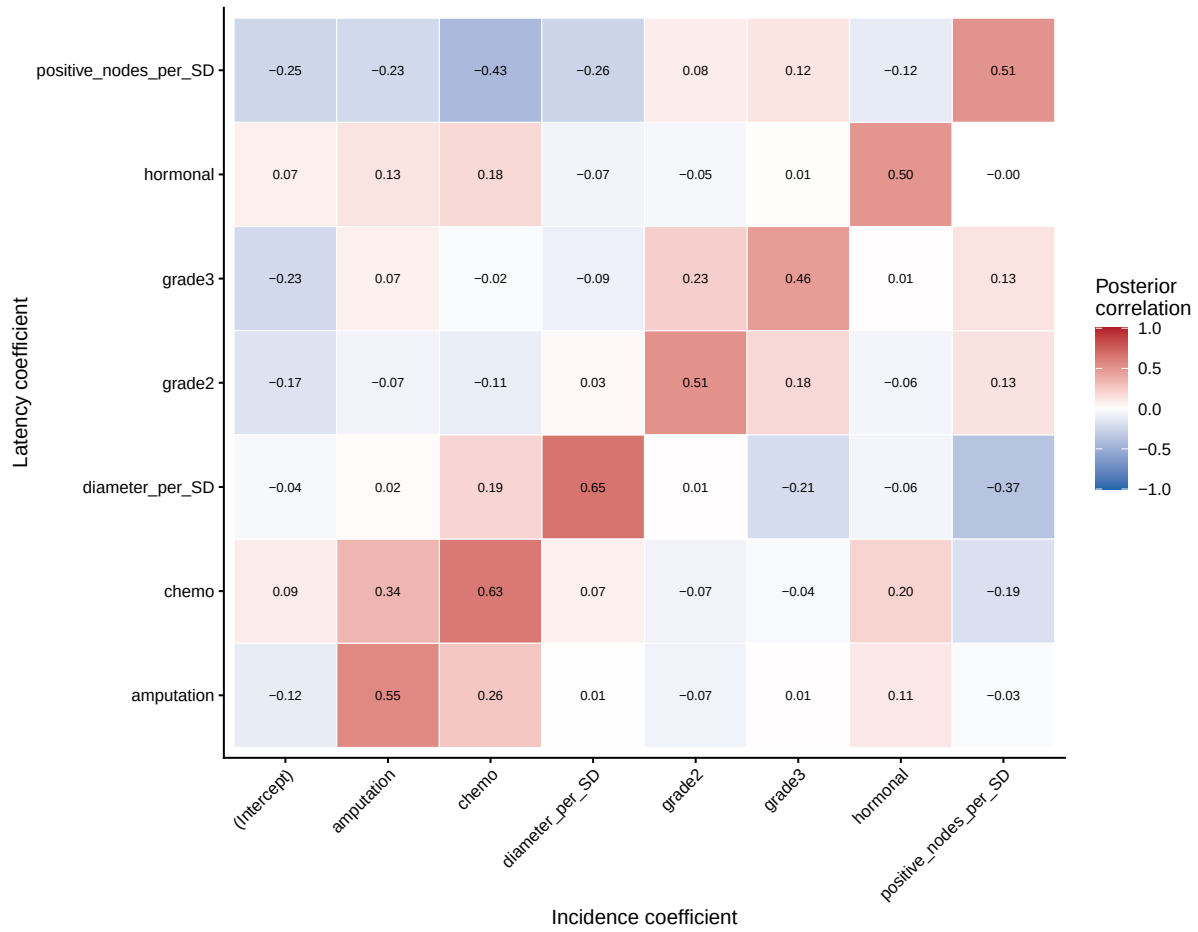


Figura 5.2: Correlações a posteriori entre os coeficientes do componente de latência (β) e os do componente de incidência (γ), sob a especificação adaptativa.

5.4.3 Comparação entre Modelos

A comparação sob a mesma verossimilhança apresentada na Tabela 5.4 mostra que a transformação adaptativa apresenta o menor DIC (952.76), enquanto seu WAIC (965.02) está apenas 0.29 acima do menor valor obtido pelo modelo PH. O modelo adaptativo também apresenta melhora em relação ao modelo paramétrico Weibull com fração de cura segundo ambos os critérios. Assim, a estimação da transformação acrescenta flexibilidade estrutural, preservando o desempenho preditivo observado entre os modelos semiparamétricos.

Tabela 5.4: Comparação dos modelos. WAIC e DIC utilizam a mesma verossimilhança baseada nos intervalos de ocorrência dos eventos para todos os modelos Bayesianos.

Modelo	WAIC	DIC
Riscos proporcionais (PH)	964.72	954.83
Chances proporcionais (PO)	966.09	952.97
Transformação adaptativa	965.02	952.76
Cura por mistura Weibull	1008.99	1006.37

O modelo semiparamétrico padrão de Cox com cura por mistura foi ajustado separadamente por meio do **smcure**, utilizando os mesmos preditores contínuos padronizados e os mesmos contrastes clínicos. Sua curva de sobrevivência marginal ajustada acompanha de perto a curva de Kaplan–Meier e os ajustes Bayesianos semiparamétricos apresentados na Figura 5.3; em relação à curva média a posteriori do modelo adaptativo, o RMSE na grade é 0.016 e a diferença absoluta máxima é 0.036. Essa concordância fornece uma referência clássica externa, enquanto o modelo adaptativo incorpora incerteza a posteriori, inferência direta para o componente de cura e uma transformação estimada a partir dos dados.

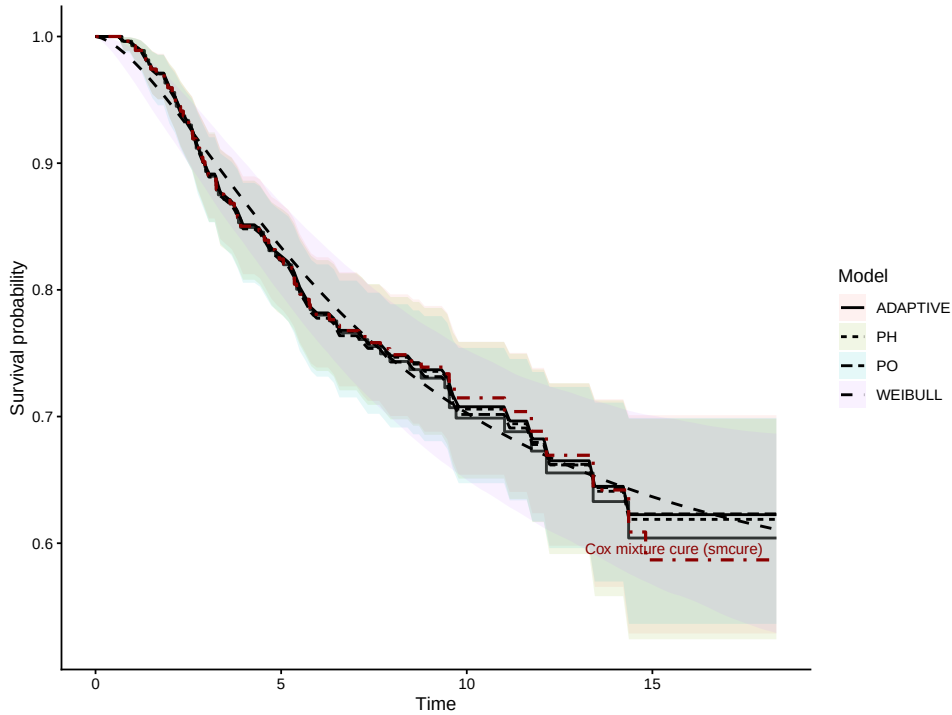


Figura 5.3: Estimativa de Kaplan–Meier e curvas ajustadas de sobrevivência marginal para os modelos adaptativo, PH, PO, Weibull e Cox padrão de mistura com fração de cura. As bandas sombreadas representam intervalos de credibilidade posterior de 95% para os ajustes Bayesianos.

5.4.4 Análise de Sensibilidade e Diagnósticos Gráficos

A análise de sensibilidade às priors indica que os resultados se mantêm estáveis sob as hiperpriors alternativas $\text{Gamma}(0.1, 0.1)$. Considerando todos os coeficientes de latência, incidência e transformação, a maior alteração na média a posteriori foi de 0.170 desvios-padrão a posteriori combinados. Para $\Lambda_0(t)$, o RMSE na grade entre as médias a posteriori foi 0.037 e a diferença absoluta máxima foi 0.311; para $G(x)$, os valores correspondentes foram 0.294 e 0.767 em um intervalo da média a posteriori que se estendeu até cerca de 38. Além disso, as bandas de credibilidade de 95% das duas especificações de priori apresentaram sobreposição em todos os 77 tempos de evento e em todos os 151 pontos da grade da transformação. As precisões de suavização responderam à hiperpriori conforme esperado—as médias a posteriori de τ_ψ e τ_θ passaram de 8.646 e 1.253 para 48.184 e 3.427—enquanto os efeitos de regressão e as funções estimadas permaneceram próximos entre as duas especificações. O ajuste alternativo também apresentou diagnósticos adequados para as quatro cadeias: o maior valor de $\hat{R}$ foi 1.010 e o ESS mínimo para a região central foi 485 entre todos os blocos apresentados.

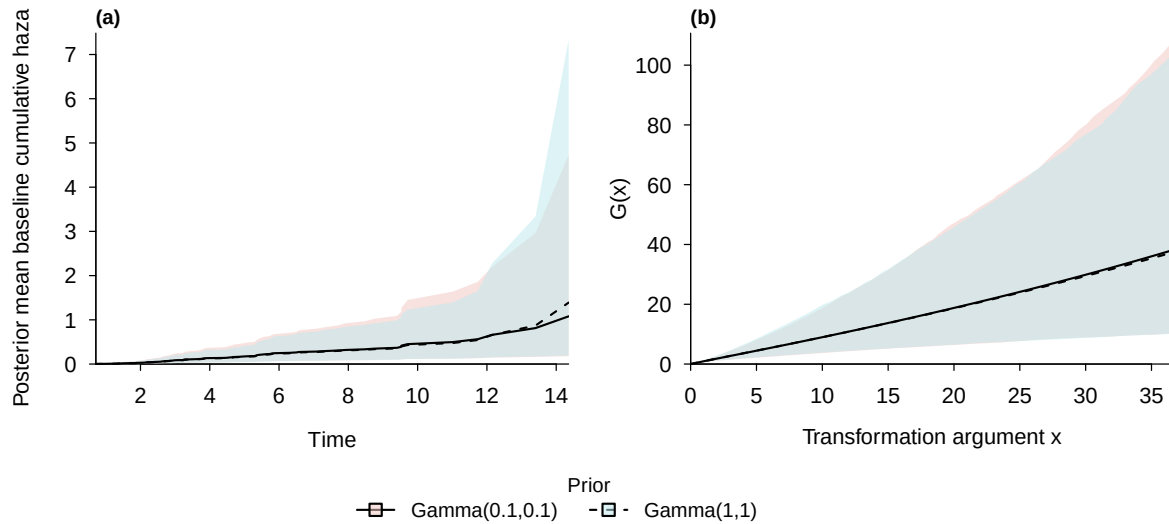


Figura 5.4: Sensibilidade aos valores das distribuições a priori sob hipera-priors de precisão de suavização $\text{Gamma}(1, 1)$ e $\text{Gamma}(0.1, 0.1)$: médias posteriores e bandas de credibilidade de 95

Para finalizar a análise, as Figuras 5.5–5.14 apresentam os *traceplots* das quatro cadeias executadas em paralelo para cada especificação, considerando os coeficientes de latência, de incidência e, no modelo adaptativo, da transformação. Em todos os painéis, as cadeias exibem o padrão característico de convergência, com oscilação rápida em torno de um nível estável, ausência de tendência sistemática ao longo das iterações e sobreposição entre as quatro trajetórias, sem que nenhuma delas permaneça em região distinta das demais. Não se observam degraus, deslocamentos abruptos de nível ou períodos prolongados de estagnação, de modo que não há indícios de falha de convergência em nenhum dos parâmetros monitorados. No bloco de latência, a mistura é satisfatória para todos os coeficientes, com exceção parcial do efeito associado ao grau tumoral mais elevado nas especificações de chances proporcionais e adaptativa, cujas trajetórias apresentam excursões mais prolongadas. Esse comportamento indica autocorrelação mais alta e, conseqüentemente, tamanho efetivo de amostra menor para esse parâmetro, embora as cadeias permaneçam concentradas na mesma região do espaço paramétrico. No bloco de incidência, as cadeias também se mostram estacionárias e bem misturadas, porém com trajetórias marcadas por excursões assimétricas, compatíveis com distribuições a posteriori assimétricas. Cabe destacar que a adequação da mistura não deve ser confundida com identificabilidade: a amplitude das trajetórias aproxima-se daquela induzida pelas distribuições a priori, o que é coerente com a limitada informação amostral disponível para esse componente, já evidenciada pelos desvios-padrão a posteriori próximos aos das priors. Os coeficientes da transformação adaptativa apresentam mistura rápida, com trajetórias densas e centradas em valores próximos de zero. Como o vetor nulo corresponde à transformação identidade, esse padrão reforça a conclusão de que a forma estimada permaneceu próxima da estrutura de riscos proporcionais. Por fim, a comparação entre o modelo adaptativo sob a especificação original e sob a hiperpriori $\text{Gama}(0,1; 0,1)$ revela trajetórias visualmente equivalentes, com níveis e amplitudes semelhantes em todos os parâmetros, sugerindo que as conclusões obtidas não são sensíveis à escolha dessa hiperpriori.

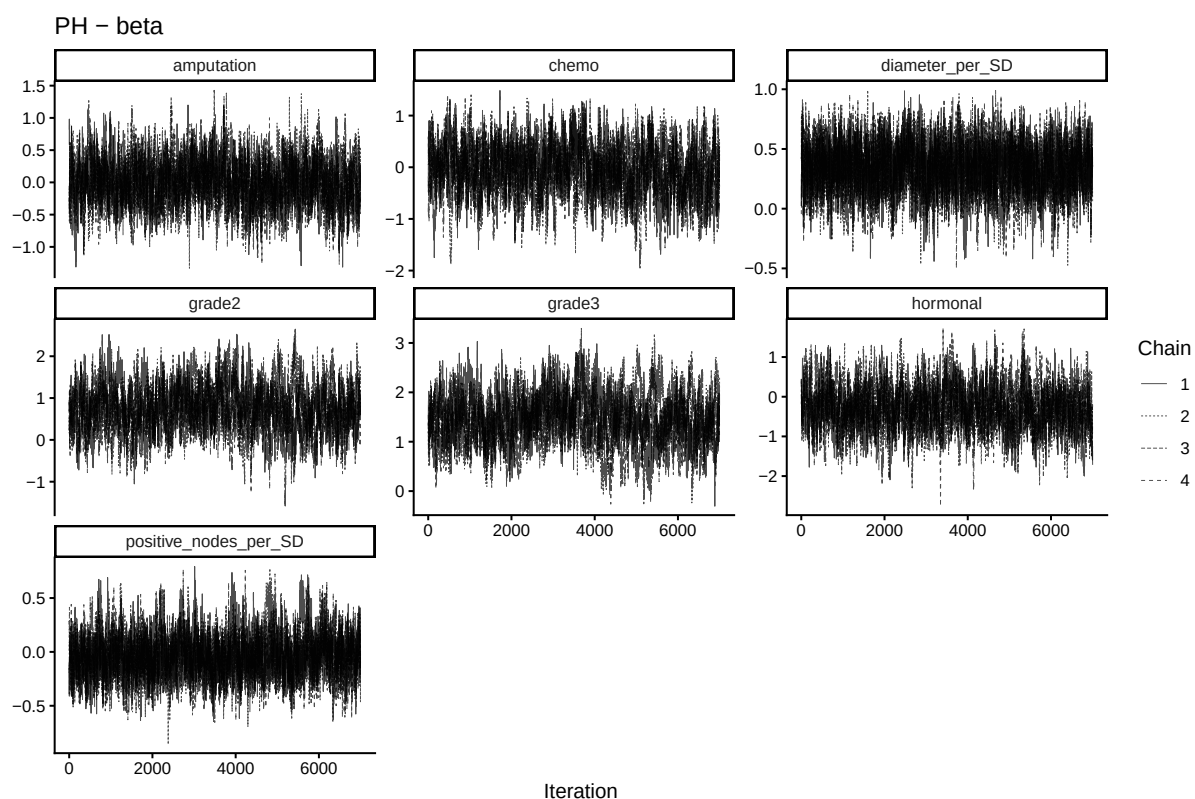


Figura 5.5: *Traceplots* das quatro cadeias para os coeficientes de latência β no modelo de riscos proporcionais (RP).

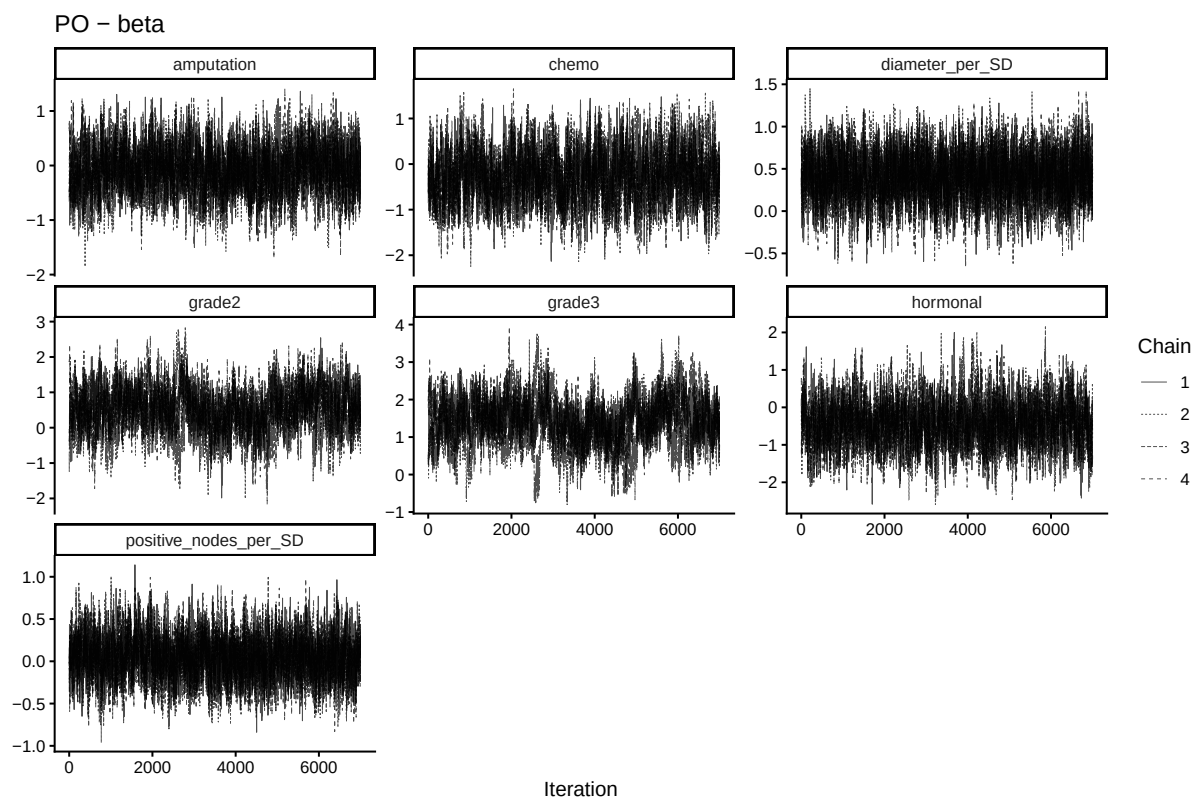


Figura 5.6: *Traceplots* das quatro cadeias para os coeficientes de latência β no modelo de chances proporcionais (CP).

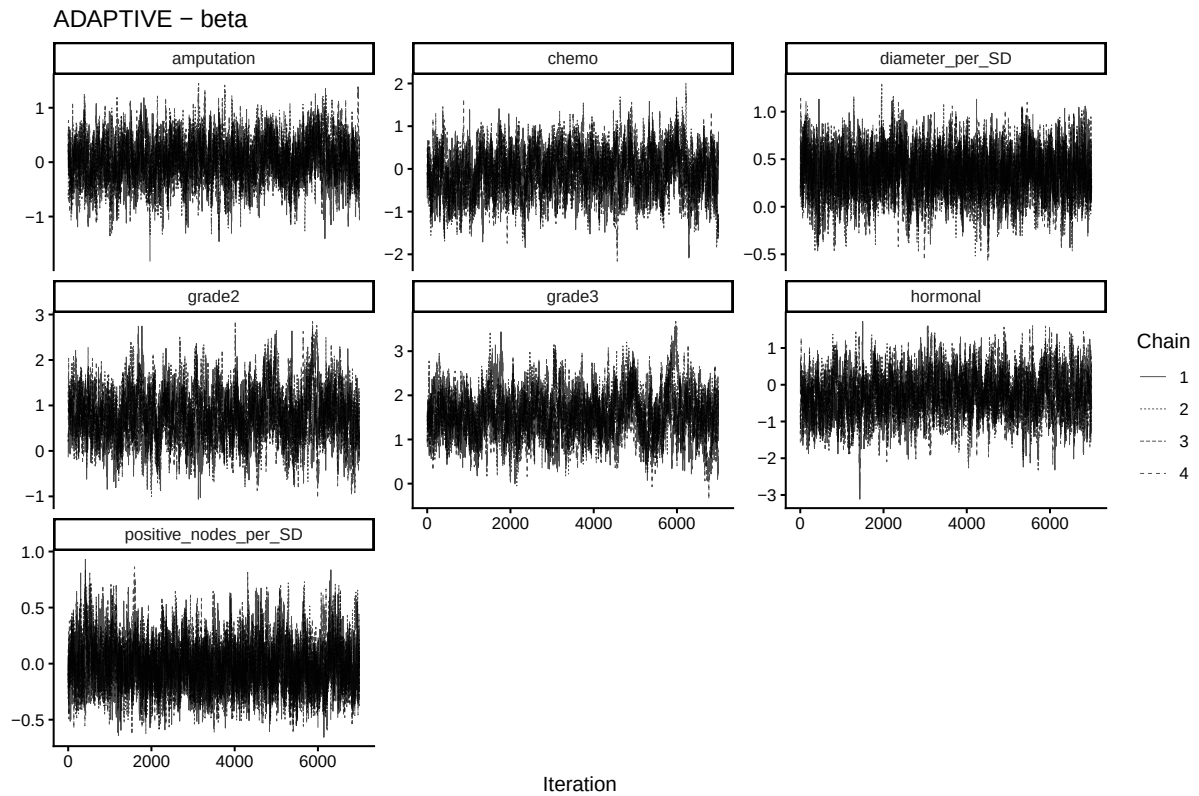


Figura 5.7: *Traceplots* das quatro cadeias para os coeficientes de latência β no modelo com transformação adaptativa.

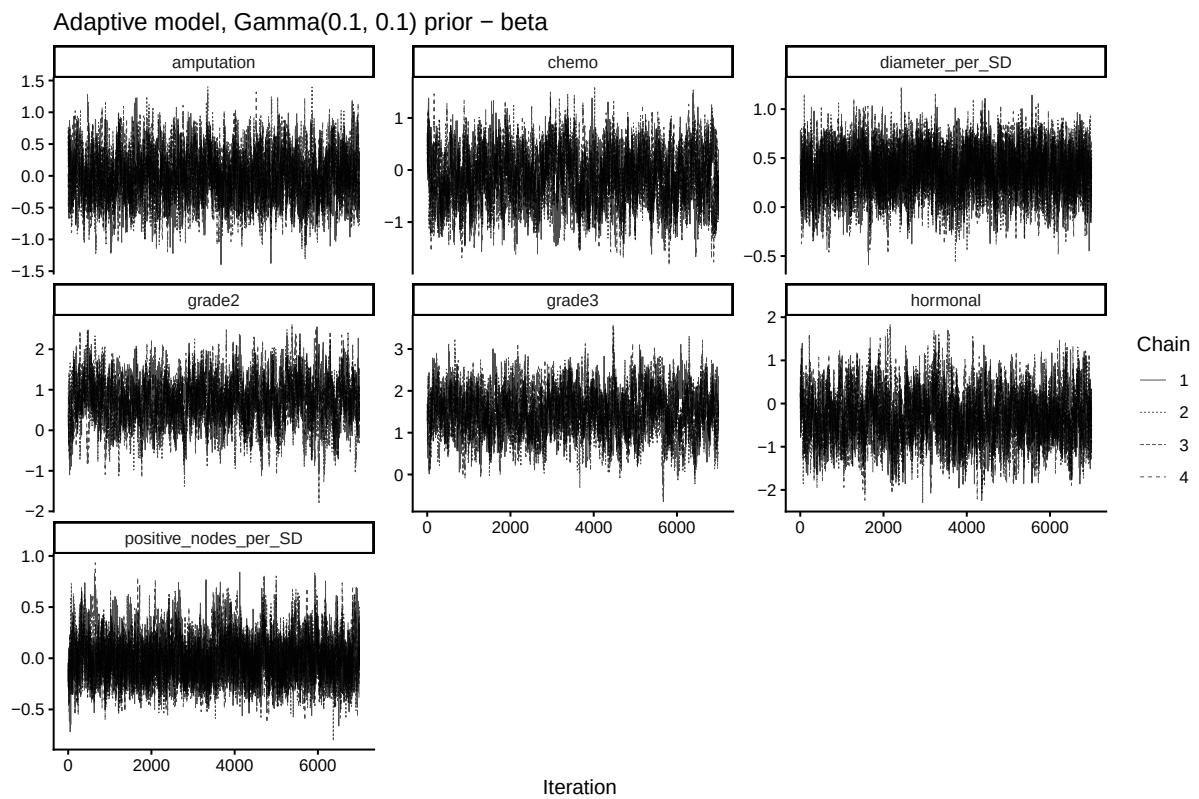


Figura 5.8: *Traceplots* das quatro cadeias para os coeficientes de latência β no modelo com transformação adaptativa sob a hiperpriori Gama(0,1; 0,1).

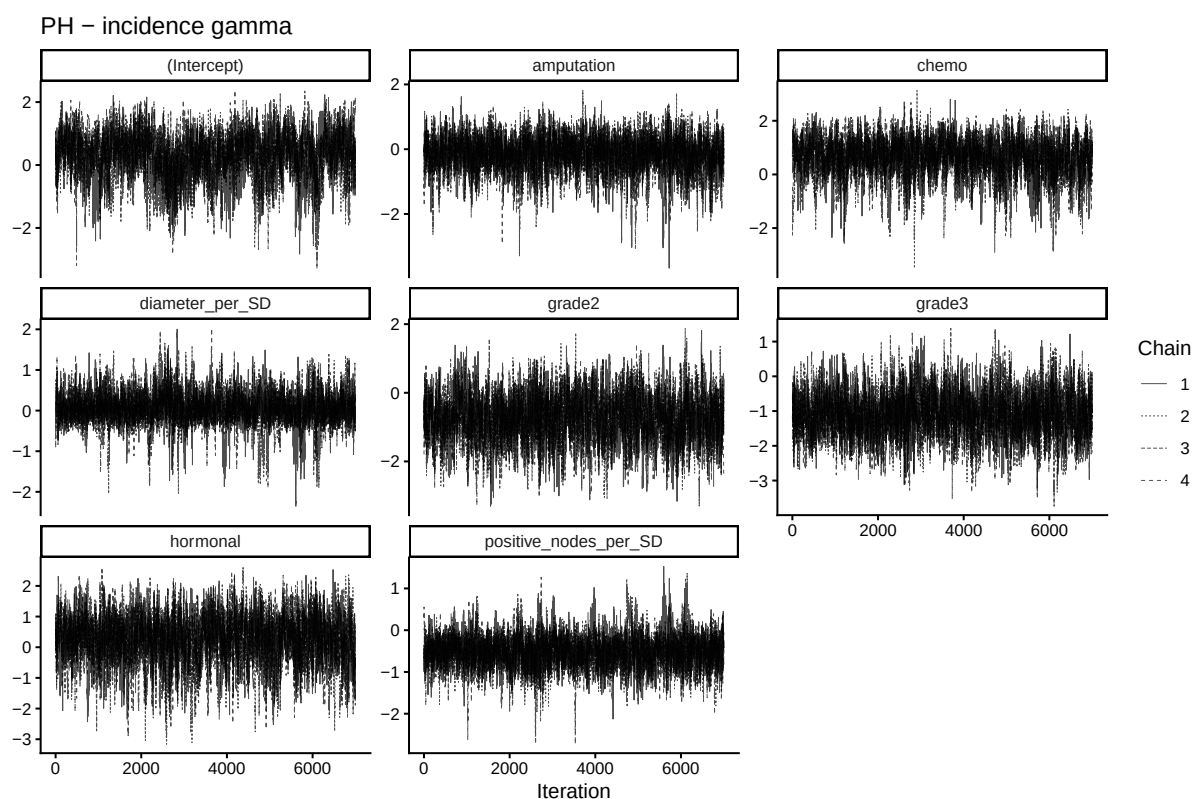


Figura 5.9: *Traceplots* das quatro cadeias para os coeficientes de incidência γ no modelo de riscos proporcionais (RP).

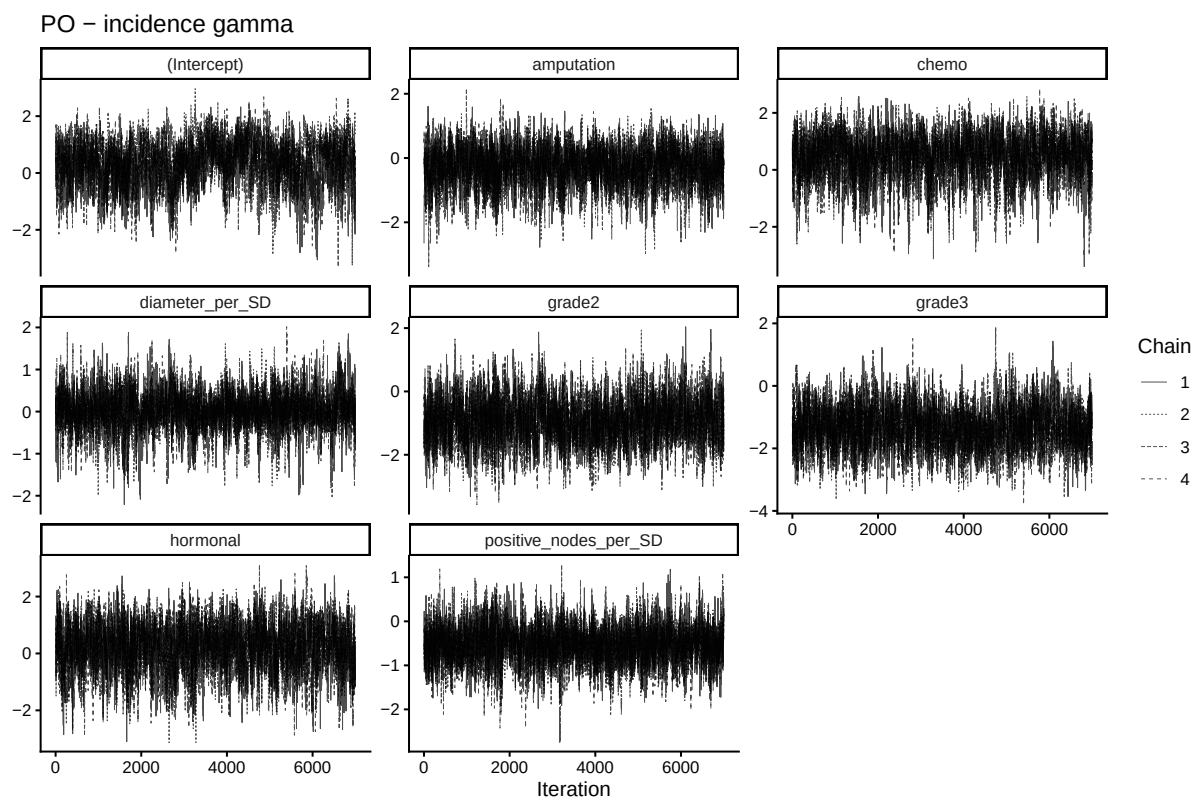


Figura 5.10: *Traceplots* das quatro cadeias para os coeficientes de incidência γ no modelo de chances proporcionais (CP).

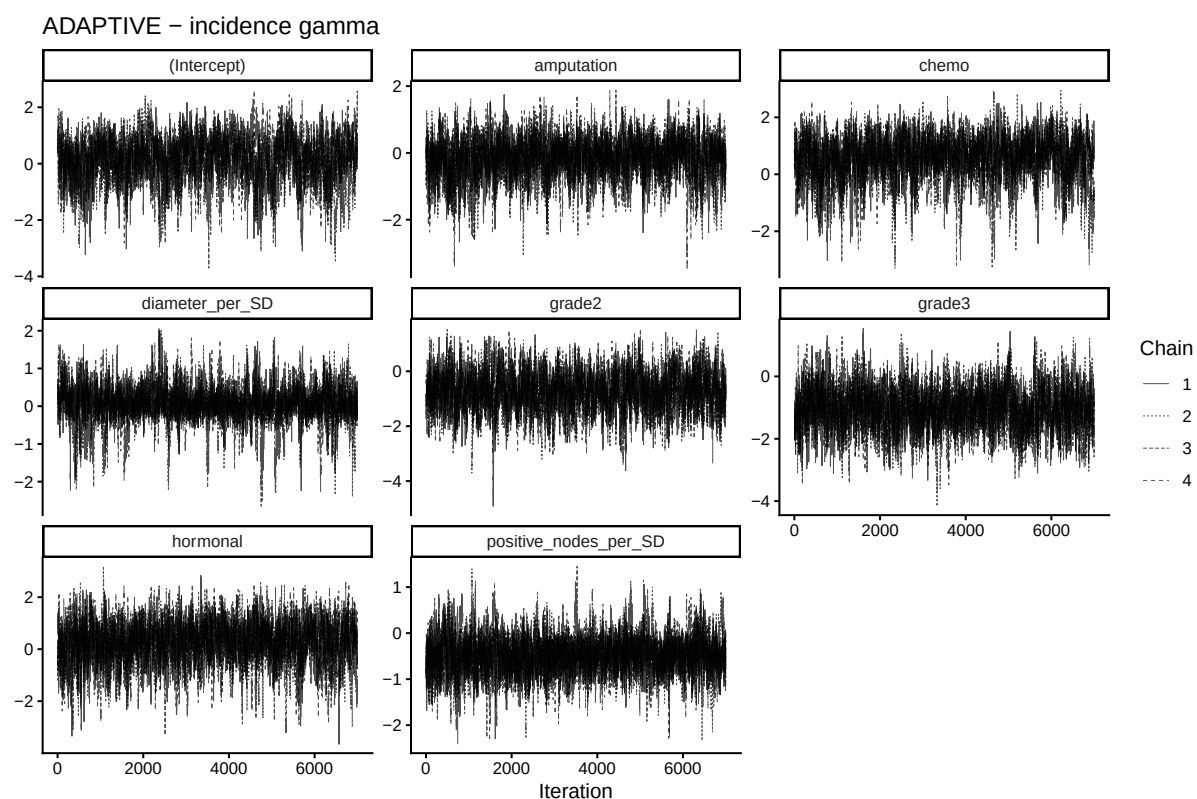


Figura 5.11: *Traceplots* das quatro cadeias para os coeficientes de incidência γ no modelo com transformação adaptativa.

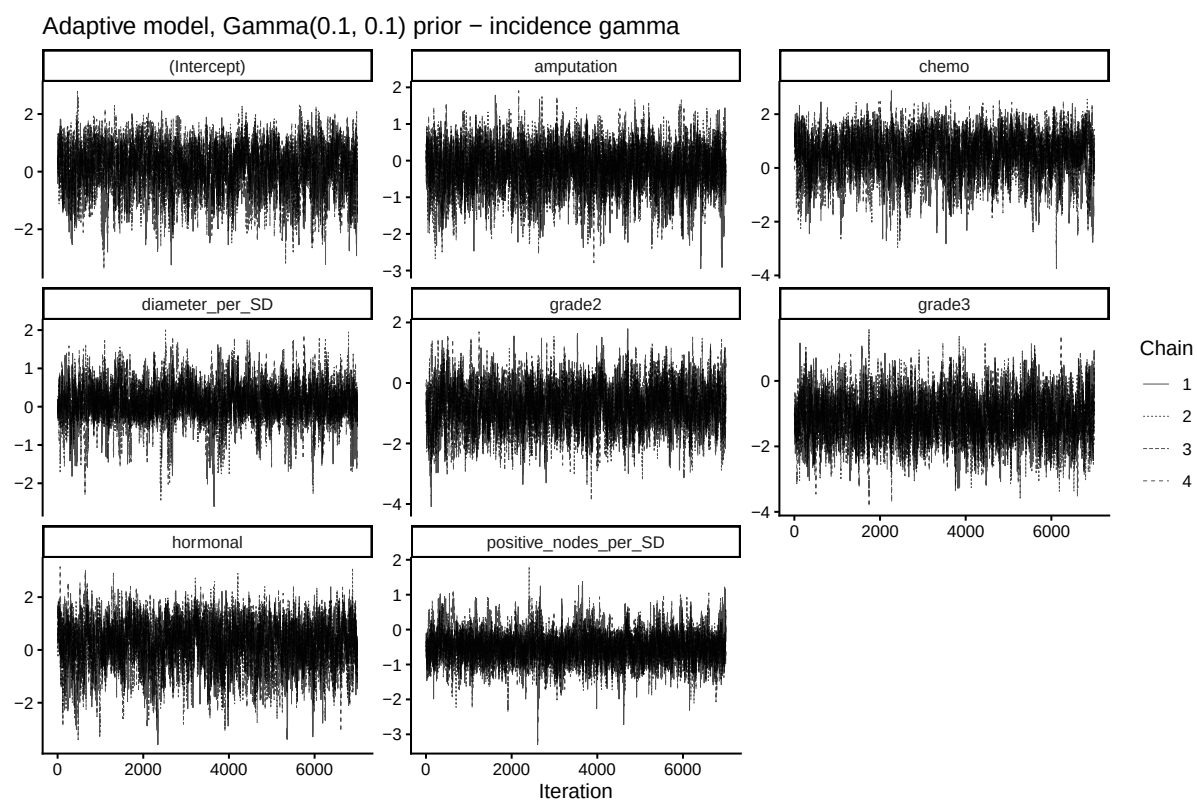


Figura 5.12: *Traceplots* das quatro cadeias para os coeficientes de incidência γ no modelo com transformação adaptativa sob a hiperpriori Gama(0,1; 0,1).

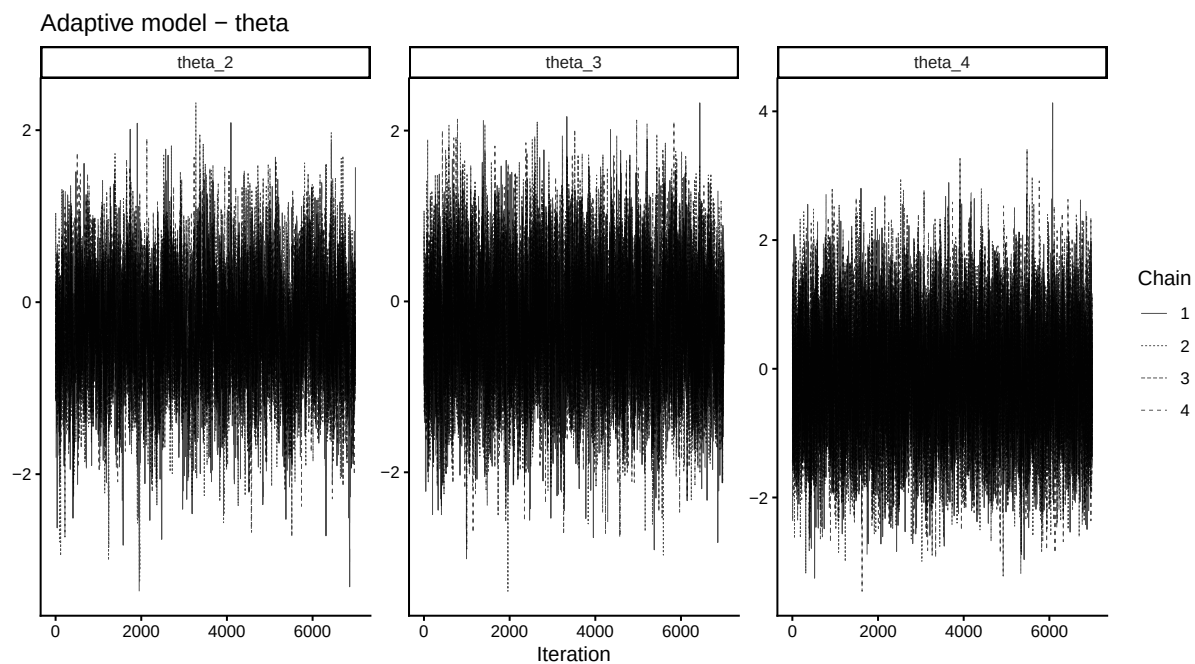


Figura 5.13: *Traceplots* das quatro cadeias para os coeficientes de transformação θ no modelo com transformação adaptativa.

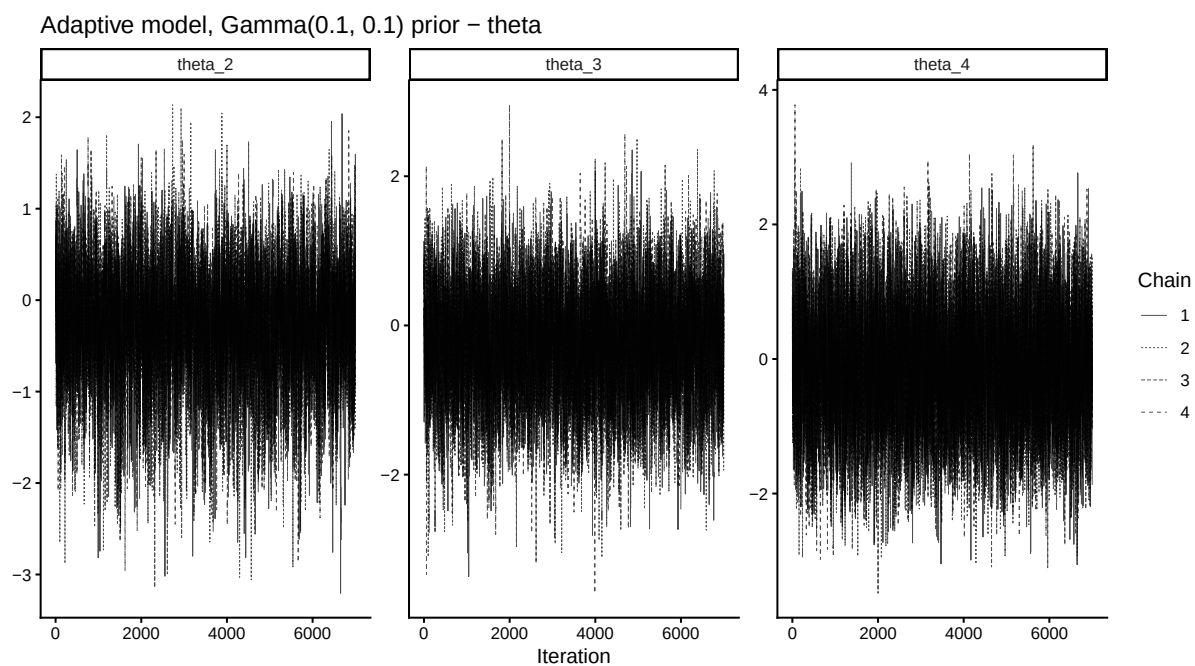


Figura 5.14: *Traceplots* das quatro cadeias para os coeficientes de transformação θ no modelo com transformação adaptativa sob a hiperpriori Gama(0,1; 0,1).

Conclusão do trabalho

Este trabalho apresentou um *framework* Bayesiano semiparamétrico penalizado de transformação do risco acumulado para dados de sobrevivência com fração de cura. A proposta integra três componentes: um modelo de transformação do risco acumulado, uma formulação de cura por mistura que separa os componentes de incidência e latência e uma especificação semiparamétrica para o risco acumulado *baseline*. A transformação permite representar estruturas que incluem os modelos de riscos proporcionais e de chances proporcionais como casos particulares, enquanto o componente de cura distingue a probabilidade de sobrevivência de longo prazo da distribuição do tempo até o evento entre os indivíduos suscetíveis. O risco acumulado *baseline*, por sua vez, é representado por incrementos positivos nos tempos de evento e regularizado por uma penalização de passeio aleatório de segunda ordem na escala logarítmica. Essa construção permite estimar, em um mesmo modelo, os efeitos associados à incidência, à latência, à forma da transformação e à trajetória do risco acumulado *baseline*.

Além das funções de transformação previamente especificadas, foi proposta uma transformação monótona adaptativa aos dados, construída por meio de uma spline exponencial integrada. A condição de positividade da derivada garante a monotonicidade da transformação, enquanto as restrições $G'(0) = 1$, $\theta_1 = 0$ e a ausência de intercepto no componente de latência estabelecem a identificação da escala em relação ao risco acumulado *baseline*. A penalização RW2 aplicada aos coeficientes da transformação controla a variação entre coeficientes adjacentes, enquanto uma penalização da mesma classe aplicada aos logaritmos dos incrementos do risco *baseline* regulariza sua trajetória. A inferência a posteriori foi conduzida por meio de um algoritmo Metropolis-within-Gibbs, combinando propostas de passeio aleatório multivariadas, atualizações MALA preconditionadas para os incrementos *baseline* e etapas de Gibbs para os parâmetros de precisão.

O estudo de simulação permitiu avaliar o comportamento do modelo adaptativo em diferentes tamanhos amostrais e níveis de censura. Nos cenários considerados, todas as replicações produziram ajustes válidos e o mecanismo de censura reproduziu os níveis estabelecidos no processo gerador dos dados. Os coeficientes de latência apresentaram baixos valores de viés absoluto, acompanhados por redução do RMSE com o aumento do tamanho amostral. A estimação dos componentes de incidência e transformação apresentou maior incerteza, sobretudo em alguns cenários com amostras maiores, refletindo a dificuldade de separar os efeitos associados à probabilidade de cura daqueles relacionados ao tempo de sobrevivência dos indivíduos suscetíveis. Para os coeficientes da transformação, a contração observada confirma que sua interpretação isolada é menos informativa do que a análise da função $G(\cdot)$ estimada. No nível funcional, a transformação foi recuperada com RMSE normalizado entre 12.3% e 18.5% na grade utilizada, enquanto a penalização

aplicada ao risco *baseline* se adaptou ao aumento do número de tempos de evento. Os diagnósticos de aceitação, ESS e estabilidade das cadeias também indicaram que o procedimento computacional permaneceu operacional nos cenários avaliados, mesmo quando a razão entre o número de parâmetros e o tamanho amostral foi elevada.

Na aplicação aos dados de câncer de mama do Netherlands Cancer Institute, os resultados mostraram a utilidade da decomposição entre incidência e latência e da estimação da transformação a partir dos dados. O modelo adaptativo apresentou o menor DIC entre os modelos Bayesianos avaliados e um WAIC próximo ao menor valor obtido pelo modelo de riscos proporcionais. A curva de sobrevivência marginal estimada também apresentou concordância com o ajuste semiparamétrico de Cox com cura por mistura obtido pelo *smcure*. No componente de latência, o grau tumoral III foi associado à ocorrência mais precoce de eventos entre as pacientes suscetíveis, resultado que também foi expresso por meio de contrastes de sobrevivência em 5 e 10 anos. A análise das correlações a posteriori mostrou dependência entre alguns coeficientes de incidência e latência, com correlações em valor absoluto inferiores a 0.65, indicando que parte da informação fornecida pelos dados é compartilhada entre esses dois componentes. Esse resultado, em conjunto com os desvios-padrão a posteriori do componente de incidência, evidencia a dificuldade de identificação desse componente em relação à latência. A análise de sensibilidade às priors mostrou que as alterações nas hiperpriors de suavização produziram mudanças nas precisões τ_ψ e τ_θ , mas mantiveram próximos os efeitos de regressão, o risco acumulado *baseline* e a função de transformação estimada. Os diagnósticos das quatro cadeias também apresentaram valores de $\hat{R}$ e ESS compatíveis com a estabilidade da amostragem a posteriori.

Os resultados obtidos também indicam aspectos que devem ser considerados na utilização dessa classe de modelos. A separação entre incidência e latência pode apresentar identificação limitada quando os mesmos preditores são incluídos nos dois componentes, em especial em amostras moderadas ou quando a informação sobre sobrevivência de longo prazo é reduzida. Além disso, os coeficientes individuais da transformação dependem da base spline e da regularização adotada, de modo que a interpretação deve privilegiar a função $G(\cdot)$ estimada, os contrastes de sobrevivência e outras quantidades derivadas do modelo. A dimensão do componente *baseline* também cresce com o número de tempos de evento distintos, tornando a regularização e a estratégia computacional partes relevantes da especificação do modelo.

Como extensões, a formulação geral baseada em causas competitivas permite substituir o modelo de cura por mistura por estruturas de incidência baseadas em tempo de promoção, binomial negativa ou distribuição geométrica, mantendo os componentes Bayesianos semiparamétricos associados à transformação e ao risco acumulado *baseline*. Também podem ser considerados conjuntos distintos de covariáveis para incidência e latência, bem como priors de contração para seleção ou regularização dos efeitos de regressão. Para o componente de transformação, estudos futuros podem avaliar diferentes dimensões das bases spline, posicionamentos de nós e estruturas de penalização, além de critérios preditivos baseados em validação cruzada. Outras extensões incluem modelos para dados agrupados, estruturas multiestados e modelos conjuntos para dados longitudinais e de sobrevivência com fração de cura. Essas direções ampliam a classe desenvolvida neste trabalho e permitem sua aplicação a diferentes estruturas de dados de sobrevivência.

Referências

- Andersson, T. M., Dickman, P. W., Eloranta, S., et al. (2011). Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models. *BMC Med Res Methodol*, 11:96.
- Bradburn, M. J., Clark, T. G., Love, S. B., and Altman, D. G. (2003). Survival analysis part ii: multivariate data analysis—an introduction to concepts and methods. *British journal of cancer*, 89(3):431–436.
- Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6):394–424.
- Cardoso, F. et al. (2018). 4th eso-esmo international consensus guidelines for advanced breast cancer (abc 4). *Annals of Oncology*, (12):1634–1657. Acesso em: 04 out. 2024.
- Chen, C.-M. and Lu, T.-F. C. (2012). Marginal analysis of multivariate failure time data with a surviving fraction based on semiparametric transformation cure models. *Computational Statistics & Data Analysis*, 56(3):645–655.
- Chow, S.-C. and Chang, M. (2008). Adaptive design methods in clinical trials—a review. *Orphanet journal of rare diseases*, 3(1):11.
- Collett, D. (2015). *Modelling survival data in medical research*. CRC press.
- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.
- Demarqui, F. N., Mayrink, V. D., and Ghosh, S. K. (2019). An unified semiparametric approach to model lifetime data with crossing survival curves. *arXiv preprint arXiv:1910.04475*.
- Ferlay, J., Ervik, M., Lam, F., et al. (2024). Global cancer observatory: Cancer today (version 1.0). Accessed on January 27, 2025.
- Foulkes, W. D. (2008). Inherited susceptibility to common cancers. *New England Journal of Medicine*, 359(20):2143–2153. Acesso em: 04 out. 2024.
- Giolo, S. R. and Colosimo, E. A. (2006). *Análise de sobrevivência aplicada*. Edgard Blucher.
- Goldman, L. and Schafer, A. I. (2011). *Goldman’s Cecil Medicine*. Elsevier Inc., United States of America, 24th edition.
- Greenwood, M. (1926). A report on the natural duration of cancer. Technical Report 33, Ministry of Health, London.

- Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242.
- Hartung, D. M. and Touchette, D. (2009). Overview of clinical research design. *American Journal of Health-System Pharmacy*, 66(4):398–408.
- Hunk, K. K., Robertson, J. F. R., and Bland, K. I. (2015). The breast. In B., F. C., editor, *Schwartz’s Principles of Surgery*, pages 517–520. McGraw-Hill, Texas, 10th edition.
- Instituto Nacional de Câncer (2025). *Controle do câncer de mama no Brasil: dados e números 2025*. INCA, Rio de Janeiro.
- Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.
- Kasper, D. L., Fauci, A. S., Stephen, H., et al. (2016). *Harrison’s Principles of Internal Medicine*. McGraw-Hill, 19th edition.
- Kiani, A. K., Naureen, Z., Pheby, D., Henahan, G., Brown, R., Sieving, P., Sykora, P., Marks, R., Falsini, B., Capodicasa, N., et al. (2022). Methodology for clinical research. *Journal of preventive medicine and hygiene*, 63(2 Suppl 3):E267.
- Klein, J. P. and Moeschberger, M. L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer-Verlag, New York.
- Leboo, I. (2023). Introduction to survival analysis. *Medium article*, accessed 14 October 2025.
- Lee, E. T. and Wang, J. W. (2003). *Statistical methods for survival data analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience, Hoboken, NJ, third edition.
- Li, J. and Ma, S. (2013). *Survival analysis in medicine and genetics*. CRC Press.
- Lilienfeld, A. and Lilienfeld, D. (1980). *Foundations of Epidemiology*. Oxford University Press, New York.
- Maller, R. A. and Zhou, X. (1996). *Survival analysis with long-term survivors*. Wiley, Chichester.
- Mariotto, A. B., Noone, A. M., Howlader, N., Cho, H., Keel, G. E., Garshell, J., et al. (2014). Cancer survival: an overview of measures, uses, and interpretation. *J Natl Cancer Inst Monogr*, 2014(49):145–186.
- McPherson, K., Steel, C. M., and Dixon, J. M. (2000). Abc of breast diseases: Breast cancer—epidemiology, risk factors, and genetics. *BMJ*, 321(7261):624–628.
- Meeker, W. Q. and Escobar, L. A. (1998). *Statistical Methods for Reliability Data*. John Wiley & Sons, New York.
- Othus, M., Barlogie, B., LeBlanc, M. L., and Crowley, J. J. (2012). Cure models as a useful statistical tool for analyzing survival. *Clinical Cancer Research*, 18(14):3731–3736.
- Peng, Y. (2003). Fitting semiparametric cure models. *Computational statistics & data analysis*, 41(3-4):481–490.

- Ramires, T. G., Hens, N., Cordeiro, G. M., and Ortega, E. M. (2018). Estimating non-linear effects in the presence of cure fraction using a semi-parametric regression model. *Computational Statistics*, 33:709–730.
- Roberts, G. O. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268.
- Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363.
- Rodrigues, J., Cancho, V. G., de Castro, M., and Louzada-Neto, F. (2009). On the unification of long-term survival models. *Statistics & Probability Letters*, 79(6):753–759.
- Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC, Boca Raton.
- Schober, P. and Vetter, T. R. (2018). Survival analysis and interpretation of time-to-event data: the tortoise and the hare. *Anesthesia & Analgesia*, 127(3):792–798.
- Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1):239–241.
- Sociedade Americana do Câncer (American Cancer Society) (2025). Tipos de câncer de mama. Sociedade Americana do Câncer. Acesso em: 21 out. 2025.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.
- Tsodikov, A. D., Ibrahim, J. G., and Yakovlev, A. Y. (2003). Estimating cure rates from survival data: an alternative to two-component mixture models. *Journal of the American Statistical Association*, 98(464):1063–1078.
- Vahidpour, M. (2016). Cure rate models. Master’s thesis, École Polytechnique de Montréal, Montréal, Canada.
- van de Vijver, M. J., He, Y. D., van ’t Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25):1999–2009.
- Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC. *Bayesian Analysis*, 16(2):667–718.
- Wang, C. (2008). Non-parametric maximum likelihood estimation for cox regression with subject-specific measurement error. *Scandinavian Journal of Statistics*, 35(4):613–628.
- Wang, X. and Kattan, M. W. (2020). Cohort studies: design, analysis, and reporting. *Chest*, 158(1):S72–S78.
- Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594.

- Yang, L., Song, H., Peng, Y., and Tu, D. (2021). Joint analysis of longitudinal measurements and survival times with a cure fraction based on partly linear mixed and semiparametric cure models. *Pharmaceutical Statistics*, 20(2):362–374.
- Zabor, E. C., Kaizer, A. M., and Hobbs, B. P. (2020). Randomized controlled trials. *Chest*, 158(1):S79–S87.
- Zeng, D. and Lin, D. Y. (2006). Efficient estimation of semiparametric transformation models for counting processes. *Biometrika*, 93(3):627–640.