Universidade Federal da Bahia - UFBA
Instituto de Matemática e Estatística - IME
Programa de Pós-Graduação em Matemática - PGMAT
Dissertação de Mestrado

Estimação bayesiana nos modelos com respostas distais em análise de sobrevivência

Marcos Aurélio Eustorgio Filho

Área de Concentração: Estatística

Salvador - Bahia
Outubro de 2024

Estimação bayesiana nos modelos com respostas distais em análise de sobrevivência

Marcos Aurélio Eustorgio Filho

Dissertação de Mestrado apresentada ao Colegiado do Programa da Pós-Graduação em Matemática da Universidade Federal da Bahia (UFBa), como parte dos requisitos para obtenção do título de Mestre em Matemática. Área de concentração: Estatística.

Orientadora: Profa. Dra. Leila Denise Alves Ferreira Amorim
Coorientadora: Profa. Dra. Lilia Carolina Carneiro da Costa

Salvador - Bahia
Outubro 2024

Ficha catalográfica elaborada pela Biblioteca Universitária de Ciências e Tecnologias Prof. Omar Catunda, SIBI – UFBA.

E91 Eustorgio Filho, Marcos Aurélio

Estimação bayesiana nos modelos com respostas distais em análise de sobrevivência. / Marcos Aurélio Eustorgio Filho. – Salvador, 2024.

139 p.

Orientadora: Prof.ª Dr.ª Leila Denise Alves Ferreira Amorim.

Coorientadora: Prof.ª Dr.ª Lilia Carolina Carneiro da Costa.

Dissertação (Mestrado) – Universidade Federal da Bahia, Instituto de Matemática e Estatística, 2024.

1. Métodos Bayesianos. 2. Análise de Classes Latentes. 3. Modelos com Respostas Distais. 4. Análise de Sobrevivência. I. Amorim, Leila Denise Alves Ferreira. II. Costa, Lilia Carolina Carneiro da. III. Universidade Federal da Bahia. IV. Título.

CDU: 519.24

Estimação Bayesiana nos Modelos com Respostas Distais em Análise de Sobrevivência

Marcos Aurélio Eustorgio Filho

Dissertação apresentada ao Colegiado do Curso de Pós-graduação em Matemática da Universidade Federal da Bahia, como requisito parcial para obtenção do Título de Mestre em Matemática.

Banca examinadora

Leila Denise Alves Ferreira Amorim

Profª Drª Leila Denise Alves Ferreira Amorim (orientadora - UFBA)

Jony Arrais Pinto Junior

Prof°. Dr°. Jony Arrais Pinto Junior (Externo)

Rosemeire Leovigildo Fiaccone

Profa. Dra. Rosemeire Leovigildo Fiaccone (Interna)

Dedico esta conquista a minha família, amigos, e aos meus antepassados cujos caminhos foram marcados pela luta, mas que, apesar das adversidades, pavimentaram o futuro que hoje posso trilhar com mais dignidade e esperança.

Agradecimentos

Agradeço primeiramente ao Criador de todas as coisas por toda força, discernimento e oportunidades apresentadas, e que foi fundamental para minha jornada até este momento, me dando a força necessária para superar todos os obstáculos. Agradeço especialmente aos meus familiares, minha mãe Helenilza, meu padrinho Fernando e minha tia Rita, por todo suporte, cuidado, carinho e atenção, que foram fundamentais ao longo da minha vida e da minha jornada acadêmica.

Gostaria de expressar minha profunda gratidão aos meus amigos Igor e Fernando, e à minha amiga Tamilis, que foram fundamentais durante minha jornada desde a graduação. Eles me auxiliaram com sugestões, no processo de coleta de dados, e ofereceram apoio incondicional em momentos de dúvida, o que fez toda a diferença para a conclusão do meu trabalho de graduação. Cada conversa e colaboração foi essencial para o desenvolvimento e finalização deste trabalho, e sou eternamente grato pela dedicação e parceria de vocês. Agradeço imensamente à minha amiga Jéssica, minha parceira de estudos e trabalhos, e, além disso, alguém que, assim como eu, enfrentou diversos desafios para permanecer na universidade, dividindo seu tempo entre estudos, longas jornadas de trabalho, e a rotina pessoal. Obrigado por ser essa pessoa maravilhosa e guerreira, em quem sempre me inspirei e que foi uma luz em muitos momentos em que pensei que não conseguiria.

Agradeço também às minhas amigas Laila e Michele por todo o incentivo, conselhos e parceria de estudos durante este período de mestrado, no qual passamos várias tardes nos encontrando para estudar no IME. Estendo meus agradecimentos à minha amiga Rose e aos meus amigos Pedro e Renata, que sempre estiveram disponíveis para conversar e me acolher, especialmente nos momentos em que precisei ficar em Salvador. Agradeço também a Ivalbert, que, com suas discussões estatísticas inusitadas durante as muitas caronas de volta para casa, tanto me ajudou. Vocês também foram fundamentais durante este período. Continuando os agradecimentos, não posso deixar de mencionar meus amigos Matheus, Caio, Marcelo, João e minha amiga Júlia, por todos os anos de amizade desde a infância, sendo companheiros de múltiplas versões de mim, desde quando aprendi a escrever minhas primeiras palavras até este momento da dissertação.

Também gostaria de agradecer ao Departamento de Estatística da UFBA, minha *alma mater*, e a todos os seus docentes, que foram fundamentais para minha formação desde a graduação até o mestrado, oferecendo excelentes aulas, projetos e oportunidades, como as iniciações científicas das quais fiz parte em dois momentos e que foram essenciais para minha trajetória. Em especial, gostaria de agradecer às professoras Dras. Denise Viola, Leila Amorim, Rosemeire Fiaccone e Lilia Costa por suas contribuições imensuráveis durante minha passagem pela universidade. À Dra. Denise, meu muito obrigado por todos os momentos de orientação de matrícula, pelas aulas de cálculo, conselhos e acolhimento em diversas ocasiões em que me senti desanimado. À Dra. Rosemeire, agradeço imensamente pelas broncas, puxões de orelha, conselhos pessoais e profissionais, que foram e continuam sendo fundamentais na minha carreira. À Dra. Lilia, muito obrigado por todas as aulas de probabilidade super divertidas e por sua forma leve e inteligente de explicar qualquer assunto, por mais complexo que fosse, tornando-o acessível e claro. Agradeço também pela disponibilidade em tirar dúvidas e por acompanhar de perto meu desenvolvimento ao longo do mestrado.

Em especial, expresso meus profundos agradecimentos à Dra. Leila, minha orientadora desde a graduação, por todo o seu empenho com a ciência e com a estatística, por sua organização e disciplina, e por todas as excelentes aulas, que me encantaram e me fizeram desejar ser estatístico, assim como a senhora. Muito obrigado por todas as oportunidades, orientações, correções, advertências e conselhos. A senhora foi a principal responsável pela minha trajetória como profissional estatístico e a maior incentivadora do meu retorno à universidade e da conclusão deste mestrado.

Expresso também meus agradecimentos ao Centro de Integração de Dados e Conhecimentos para Saúde (CIDACS), ao professor Dr. Maurício L. Barreto e a toda equipe da plataforma CIDACS-Clima pelo apoio, compreensão e disponibilização da estrutura computacional de alto desempenho, imprescindível para a execução deste trabalho. Agradeço também a toda equipe do projeto PrEP1519, especialmente à Dra. Inês Dourado e ao Dr. Laio Magno, pela coordenação durante os anos em que atuei como estatístico no projeto e, principalmente, pela disponibilização do conjunto de dados utilizado neste trabalho.

"As causas não controladas que podem influenciar o resultado são sempre estritamente inumeráveis".

Ronald Fisher

Resumo

Modelos de respostas distais englobam metodologias para estimar o efeito de variáveis latentes sobre desfechos observados na presença de outros preditores observados, o que aumenta a complexidade matemática dos modelos. Técnicas recentes procuram estimar o efeito de classes latentes sobre desfechos distais de duas maneiras: incorporando erros de mensuração diretamente na modelagem (uma etapa) ou utilizando regras de classificação para alocar indivíduos em classes e, em seguida, tratar essas classes como preditores observados em um modelo estrutural (três etapas). Embora o método de uma etapa seja mais robusto, ele é frequentemente preterido devido à sua complexidade, que exige a re-estimação de parâmetros sempre que novas variáveis são incluídas. O método de três etapas, por sua vez, tende a subestimar os efeitos dos preditores latentes. Grande parte da pesquisa sobre a estimação de efeitos de classes latentes está focada em desfechos contínuos ou categóricos. No contexto de análise de sobrevivência, há poucas abordagens frequentistas para a estimação simultânea em modelos com desfechos distais. Este trabalho propõe uma alternativa usando inferência bayesiana para estimar efeitos de variáveis latentes em respostas do tipo tempo até o evento, permitindo a inclusão de incertezas e maior flexibilidade na estimação. A metodologia proposta foi aplicada em dados reais do projeto PrEP1519, com o intuito de estudar o efeito do risco real ao HIV no tempo até a primeira descontinuidade do tratamento preventivo ao HIV em adolescentes. Estudos de simulação foram realizados para avaliar as propriedades dos estimadores do Método Bayesiano Modal Simplificado (BSM) e do Método Bayesiano Simultâneo (BS), ambos propostos nesta dissertação para análise de respostas distais definidas por tempos de falha censurados. Os resultados dos estudos de simulação indicam que o método Bayesiano Simultâneo (BS) reduz significativamente o viés na estimação do efeito associado às classes latentes no desfecho distal. Além disto, este método também permite a inclusão de preditores observados adicionais no modelo.

Palavras-chave: Métodos Bayesianos, Análise de Classes Latentes, Modelos com Respostas Distais, Análise de Sobrevivência.

Abstract

Distal outcomes models encompass methods for estimating the impact of latent variables on observed outcomes while accounting for other predictors, increasing the mathematical complexity of the models. Recent techniques aim to assess the effect of latent classes on distant outcomes in two main ways: either by directly integrating measurement errors into the model (one-step approach) or by employing classification rules to assign individuals to categories and then considering these categories as observed predictors in a structural model (three-stage approach). Although the one-step method is more robust, it is often overlooked due to its complexity, which requires re-estimating parameters whenever new variables are included. In contrast, the three-step method tends to underestimate the effects of latent predictors. A lot of research on estimating latent class effects focuses on continuous or categorical outcomes. In survival analysis, there are only a few frequentist approaches for simultaneously estimating models with distal outcomes. This work proposes an alternative approach using Bayesian inference to estimate latent variable effects in time-to-event responses. This allows for the inclusion of uncertainties and provides greater flexibility in estimation. The proposed methodology was used to analyze real data from the PrEP1519 project. The goal was to examine how HIV real-risk behaviors impact the time until the first discontinuation of HIV preventive treatment in adolescents. We conducted simulation studies to assess the properties of the estimators from the Bayesian Simplified Modal (BSM) Method and Bayesian Simultaneous (BS) Method, both of which were proposed in this dissertation for analyzing distal outcomes defined by censored failure times. The results of the simulation studies show that the Bayesian Simultaneous (BS) method significantly reduces bias when estimating the effect of latent classes on the distal outcome. Additionally, this method allows for the inclusion of extra observed predictors in the model.

Key words: Bayesian Methods, Latent Class Analysis, Distal Outcomes, Survival Analysis.

Lista de Figuras

Lista de Tabelas

Lista de siglas

AICM Critério de Informação de Akaike Monte Carlo.

AKP Adolescentes de população-chave.

aMSM Adolescentes Homens que fazem Sexo com Homens.

aTGW Adolescentes Transexuais.

BCH Método de correção de viés desenvolvido por Bolck, Croon e Hagenaars.

BICM Critério de Informação Bayesiano Monte Carlo.

BMP Máxima Probabilidade Bayesiana.

BPC Pseudo-Classe Bayesiana.

BS Método Bayesiano Simultâneo.

BSM Método Bayesiano Modal Simplificado.

CPU Unidade Central de Processamento (Central Processing Unit, em inglês).

DBS Mancha de Sangue Seco em Papel (Dried Blood Spots, em inglês).

DIC Critério de Informação de Desvio.

DST Doenças Sexualmente Transmissíveis.

FTC Emtricitabina.

HIV Vírus da Imunodeficiência Humana.

HMC Monte Carlo Hamiltoniano.

HSH Homens que fazem Sexo com Homens.

HTLV Vírus T-Linfotrópico Humano.

IST Infecções Sexualmente Transmissíveis.

JAGS Just Another Gibbs Sampler.

LCA Análise de Classes Latentes, em inglês.

LTB Método de correção de viés desenvolvido por Lanza, Tan e Bray.

MCMC Cadeias de Markov Monte Carlo.

ML Método de correção de viés desenvolvido por Vermunt, 2010 baseado em máxima verossimilhança.

MLG Modelos lineares generalizados.

PEP Profilaxia Pós-Exposição ao HIV.

PrEP Profilaxia Pré-Exposição ao HIV.

PrEP1519 Estudo desenvolvido entre adolescentes de 15 a 19 anos para avaliar a efetividade da PrEP ao HIV.

SEM Modelos de Equações Estruturais.

STAN State-of-The-Art platform for statistical modeling and high-performance statistical computation.

TDF Fumarato de Tenofovir Desoproxila.

WinBUGS Bayesian inference Using Gibbs Sampling runs on Microsoft Windows.

Conteúdo

Capítulo 1

Introdução

Nos últimos anos tem-se verificado um elevado interesse por métodos relacionados à Modelagem com Variáveis Latentes e suas aplicações (Amorim et al., 2010), que são usados para compreender complexas inter-relações entre múltiplos fatores em estudos sofisticados. A disponibilidade de softwares estatísticos tem colaborado para seu uso. Apesar de serem métodos tradicionais nas áreas de ciências sociais e economia, observa-se a recente ampliação deste uso em outras áreas do conhecimento. Neste trabalho, o interesse está centrado em modelagem com variáveis latentes categóricas, particularmente a Análise de Classes Latentes (LCA, em inglês), que é definida com base em indicadores categóricos. Tradicionalmente extensões de LCA, com a inclusão de covariáveis, são utilizadas para entender que características podem predizer o pertencimento a uma classe latente. Quando os preditores são observados e a variável resposta (ou desfecho) é latente, os modelos matemáticos são bem entendidos (Lanza et al., 2007). No entanto, se o preditor é latente e a variável resposta é observada, tem-se um problema matemático mais complexo, que mais recentemente tem sido discutido na literatura no âmbito de predição de um desfecho distal a partir de classes latentes. Várias potenciais aplicações podem se favorecer da possibilidade deste tipo de predição, fornecendo informações etiológicas em como a confluência de características e comportamentos pode predizer um desfecho de interesse. Exemplos na literatura incluem a predição de desfechos relacionados à dor a partir de classes latentes definidas pelas barreiras encontradas para gerenciamento das dores em pacientes com câncer (Roberts e Ward, 2011); predição de depressão infantil em classes latentes relacionadas à vitimização pelos colegas/amigos (Nylund et al., 2007) ou ainda a associação entre problemas comportamentais e acadêmicos no início do ensino fundamental com atendimento psicológico, baixo desempenho e evasão escolar ao final do ensino médio (Darney et al., 2013). Vários procedimentos estatísticos têm sido propostos em LCA com desfechos distais, incluindo a regra de atribuição baseada na probabilidade máxima (Nagin, 2005), a técnica de alocação múltipla em pseudoclasses (Bandeen-Roche et al.,

1997), um método de estimação baseado no modelo LCA com desfechos distais (Lanza, Tan e Bray, 2013) e métodos com correções baseadas nas probabilidades de alocação nas classes latentes (Asparouhov e Muthén, 2014a).

Do ponto de vista matemático, a análise de classes latentes está intimamente relacionada com a área da Estatística conhecida como modelos de mistura discretos. Em sua forma mais simples, LCA pode ser vista como um modelo de análise fatorial para dados categóricos. Neste contexto, van Lang et al. (2006) definem LCA como uma técnica estatística que visa identificar o menor número de classes (ou grupos) de indivíduos com padrões similares de comportamentos que possam explicar as relações existentes em um conjunto de itens observados. O interesse geralmente não está apenas no efeito direto das características observadas durante o estudo, mas também nos fatores não mensurados ou ainda em situações em que as variáveis não podem ser mensuradas perfeitamente. Nesta abordagem, primeiramente identificam-se classes homogêneas de indivíduos, de acordo com suas respostas aos itens. Posteriormente, para cada indivíduo, é calculada a probabilidade de pertencer a uma classe específica e, para os indivíduos que pertencem a uma classe específica, a probabilidade de responder positivamente para todos os itens utilizados na análise.

No entanto, muitos pesquisadores geralmente também estão interessados nas causas e/ou consequências do pertencimento às classes latentes. Existem diferentes formas de proceder LCA com desfechos distais, sendo as duas formas mais comumente utilizadas denominadas de procedimentos em uma e em três etapas (Bakk, Tekle e Vermunt, 2013; Bolck, Croon e Hagenaars, 2004; Dayton e Macready, 1988; Hagenaars, 1990; B. Muthén, 2004; Vermunt, 2010; Yamaguchi, 2000). Quando todas as suposições do modelo são obedecidas, o procedimento mais complexo de uma etapa é melhor do ponto de vista estatístico, pois ao realizar estimação conjunta dos parâmetros é incorporado o erro de mensuração, associado à incerteza do pertencimento de uma unidade amostral à cada classe latente. Todavia, a maior parte dos pesquisadores aplicados usa o método mais simplificado, que considera as três etapas (Morin et al., 2019; Olino et al., 2010) com desfechos distais. Trabalhos recentes na literatura têm discutido novos procedimentos de estimação e iniciado uma comparação entre os mesmos para avaliação das vantagens e limitações de cada um deles (Asparouhov e Muthén, 2014a; Bakk, Tekle e Vermunt, 2013). Mais recentemente, têm sido propostos métodos de estimação bayesianos em problemas associados aos modelos de mistura com respostas distais, como o definido para respostas contínuas gaussianas (Costa, Amorim e Bispo, 2021) ou no contexto de modelos de curvas de crescimento latentes (Smid, Depaoli e Van De Schoot, 2020).

A maior parte da pesquisa envolvendo a estimação do efeito de uma classe latente no desfecho distal está restrita a desfechos contínuos ou categóricos (Asparouhov

e Muthén, 2014a; Bakk e Kuha, 2018; Bakk, Oberski e Vermunt, 2016; Bakk, Tekle e Vermunt, 2013; Bray, Lanza e Tan, 2015; Clark e Muthén, 2009; Lanza, Tan e Bray, 2013). No contexto de análise de sobrevivência, técnicas de estimação conjunta envolvendo modelos de classes latentes com funções taxas de falha fragmentadas constantes, considerando procedimentos em uma ou duas etapas, têm sido utilizadas (Asparouhov, Masyn e Muthen, 2006; Larsen, 2004; B. Muthén e Asparouhov, 2009). Lythgoe, Garcia-Fiñana e Cox (2019), por sua vez, compararam diferentes métodos em 1, 2 ou 3 etapas, incorporando o modelo de Cox. Não foram encontrados na literatura procedimentos que adotem uma abordagem bayesiana nos procedimentos de estimação em análise de sobrevivência com respostas distais.

Neste contexto, esta dissertação tem como objetivo propor procedimentos de estimação para efeitos de variáveis latentes categóricas em desfechos distais em análise de sobrevivência considerando-se uma abordagem bayesiana. Este trabalho está estruturado em sete capítulos. No Capítulo 2 é apresentada uma revisão dos conceitos fundamentais em modelagem com variáveis latentes categóricas, com foco em análise de classes latentes e suas extensões envolvendo variáveis externas, enquanto no Capítulo 3 sumariza-se o modelo de respostas distais em análise de sobrevivência, com especificação de propostas de estimação na abordagem frequentista. Também é feita revisão de detalhes a respeito da estimação do modelo de Cox na abordagem bayesiana. As propostas de estimação do efeito da variável latente categórica na resposta distal usando métodos bayesianos são apresentadas no Capítulo 4, enquanto os estudos de simulação para avaliação das propriedades dos estimadores encontram-se no Capítulo 5. Os métodos propostos são ilustrados usando dados de adolescentes de 15 a 19 anos do projeto PrEP1519 para avaliar o efeito do perfil de risco real ao HIV no tempo até a descontinuidade de tratamento preventivo (PrEP) ao vírus HIV. Informações detalhadas sobre os dados e os resultados das análises estão apresentados no Capítulo 6, enquanto considerações finais encontram-se no Capítulo 7.

Capítulo 2

Análise de Classes Latentes: Abordagens Frequentista e Bayesiana

Neste capítulo é realizada uma revisão de literatura abrangendo a análise de classes latentes, explorando as abordagens frequentista e bayesiana. Na abordagem frequentista, os parâmetros são estimados, por exemplo, pelo método da máxima verossimilhança, em que se busca encontrar os valores para os parâmetros que maximizam a função de log-verossimilhança do modelo. Além disso, são apresentadas extensões da LCA, como a LCA com variáveis externas, que permite a inclusão de variáveis observáveis, como, por exemplo, preditoras da variável latente ou de um desfecho observado que tem a variável latente como preditora.

Já a abordagem bayesiana trata os parâmetros como variáveis aleatórias, atribuindo aos mesmos distribuições probabilísticas iniciais, denominadas de distribuições *a priori*. A partir dos dados observados, é estimada à distribuição *a posteriori* dos parâmetros, levando em consideração tanto a informação dos dados quanto a informação prévia fornecida pelas distribuições *a priori*. Essa abordagem permite incorporar incertezas na estimativa dos parâmetros e realizar inferências mais robustas. Diversos métodos e algoritmos são discutidos para a estimação dos parâmetros em análise de classes latentes dentro do *framework* bayesiano. A revisão de literatura apresentada neste capítulo tem como objetivo fornecer uma visão abrangente dessas duas abordagens e suas extensões, destacando suas principais características e vantagens.

2.1 Análise de classes latentes

A análise de classes latentes (LCA, em inglês) é uma das técnicas mais conhecidas na área de modelos de mistura, e lida com situações em que existem múltiplos indicadores categóricos observados subjacentes a uma variável latente (não observada) categórica. Tal técnica é utilizada para identificar e descrever variáveis latentes categóricas definidas segundo o modelo teórico do pesquisador. Não há pressupostos de que os indicadores das classes latentes ou as classes latentes em si estão em qualquer nível de medida diferente do nominal. Como os indicadores são categóricos, sua distribuição conjunta é multinomial. Pressupostos estritos de distribuição, como a normalidade multivariada, são desnecessários (Collins e Lanza, 2009). No entanto, o modelo LCA mais usual necessita de uma suposição importante, denominada independência condicional, que especifica que os indicadores, denotados por Y, são independentes condicionais a uma classe latente C.

A LCA tem como objetivo principal identificar classes ou subpopulações de modo que as unidades observacionais, denotadas pelo índice i, sejam classificadas de acordo com seus padrões de resposta Y_{ik} para os K itens/indicadores, cada um com R_k categorias distintas resposta. Entretanto, para definir o número ideal de classes do modelo é necessário levar em consideração critérios como parcimônia, interpretabilidade e também realizar comparação de estatísticas de ajuste, afim de se encontrar o modelo mais plausível.

De acordo com Collins e Lanza (2009), Vermunt (2010) e Bakk, Oberski e Vermunt (2016), um modelo de LCA para um conjunto de K indicadores categóricos (itens) e I indivíduos, que tem como objetivo modelar a probabilidade de se obter um padrão de resposta específico para um indivíduo i, denotado por $\mathbf{Y}_i$, pode ser sumarizado através da seguinte equação:

$$P(\mathbf{Y}_i) = \sum_{c=1}^{C} P(\zeta = c) P(\mathbf{Y}_i | \zeta = c). \tag{2.1}$$

Os parâmetros a serem estimados são as prevalências das classes latentes, denotadas por $\gamma_c = P(\zeta = c)$, juntamente com as probabilidades condicionais de resposta ao item, denotadas por $\rho_{k,r_k|c} = P(Y_{ik} = r_k | \zeta = c)$ (Vermunt, 2010). A partir do pressuposto de independência dos indicadores condicional à variável latente ζ, que é conhecido como independência local, pode-se reescrever a última parte da Equação (2.1), resultando na Equação (2.2):

$$P(\mathbf{Y}_i | \zeta = c) = \prod_{k=1}^{K} P(Y_{ik} | \zeta = c) = \prod_{k=1}^{K} \prod_{r_k=1}^{R_k} \rho_{k,r_k|c}^{I(Y_{ik}=r_k)}. \tag{2.2}$$

Em outras palavras, a independência local pressupõe que, após controlar os fatores latentes que influenciam as respostas dos indivíduos, não há relação entre as respostas

de um item e as respostas em outros itens dentro de uma mesma classe latente (Bakk, Tekle e Vermunt, 2013). Esse pressuposto é fundamental para a validade e interpretação correta dos resultados da análise de classes latentes (Collins e Lanza, 2009). Além disto, na Equação (2.2) $I(Y_{ik} = r_k)$ assume valor 1 se o individuo i apresentar resposta r_k para o item k e 0 caso contrário.

Substituindo-se a Equação (2.2) em (2.1), é possível reescrever o modelo clássico de LCA com a seguinte notação:

$$P(\mathbf{Y}_i) = \sum_{c=1}^{C} \gamma_c \prod_{k=1}^{K} \prod_{r_k=1}^{R_k} \rho_{k,r_k|c}^{I(Y_{ik}=r_k)} \ . \tag{2.3}$$

Como cada indivíduo fornece apenas uma resposta à variável k, o vetor de probabilidades item-resposta para uma determinada variável, fixada a classe latente, sempre soma 1, ou seja, $\sum_{r_k=1}^{R_k} \rho_{k,r_k|c} = 1$ (Collins e Lanza, 2009).

No contexto do modelo de LCA, existem também outras probabilidades de interesse, denominadas probabilidades a posteriori ou probabilidades de classificação $P(\zeta = c|\mathbf{Y}_i)$, que permitem descrever o pertencimento a uma determinada classe latente c dado um padrão de resposta $\mathbf{Y_i}$. As estimativas das prevalências das classes latentes e das probabilidades de resposta ao item em uma LCA fornecem os elementos necessários para obter a probabilidade a posteriori de adesão através do uso do teorema de Bayes:

$$P(\zeta = c|\mathbf{Y}_i) = \frac{P(\zeta = c)P(\mathbf{Y}_i|\zeta = c)}{P(\mathbf{Y}_i)} = \frac{\gamma_c \left(\prod_{k=1}^{K} \prod_{r_k=1}^{R_k} \rho_{k,r_k|c}^{I(Y_{ik}=r_k)} \right)}{\sum_{h=1}^{C} \gamma_h \prod_{k=1}^{K} \prod_{r_k=1}^{R_k} \rho_{k,r_k|h}^{I(Y_{ik}=r_k)}} \ . \tag{2.4}$$

Considerando-se os resultados obtidos na Equação (2.4), cada indivíduo tem um vetor de probabilidades posteriores, com dimensão igual ao número de classes do modelo LCA.

2.1.1 Estimação no modelo de classes latentes

No contexto de modelos de classes latentes, usando abordagem frequentista, as estimativas são obtidas através da maximização da função de log-verossimilhança para os dados via algoritmos iterativos, pois não há forma analítica exata para obtenção das estimativas dos parâmetros. Nesse contexto a função de log-verossimilhança assume a seguinte forma:

$$\log(L) = \sum_{i=1}^{N} \log(P(\mathbf{Y}_i)) = \sum_{i=1}^{N} log \left[\sum_{c=1}^{C} \gamma_c \prod_{k=1}^{K} \prod_{r_k=1}^{R_k} \rho_{k,r_k|c}^{I(Y_{ik}=r_k)} \right], \tag{2.5}$$

que é maximizada em relação aos parâmetros γ_c e $\rho_{k,r_k|c}$, usando o algoritmo EM (*Expectation-Maximization*) (Dempster, Laird e Rubin, 1977).

Segundo Linzer e Lewis (2011), o algoritmo EM prossegue iterativamente, começando com valores iniciais arbitrários para $\hat{\gamma}_c$ e $\hat{\rho}_{k,r_k|c}$, que são nomeados como $\hat{\gamma}_c^{old}$ e $\hat{\rho}_{k,r_k|c}^{old}$. Na etapa da esperança (E) são calculadas as probabilidades de pertencimento à classe desconhecida via Equação (2.4), substituindo os parâmetros pelos respectivos valores $\hat{\gamma}_c^{old}$ e $\hat{\rho}_{k,r_k|c}^{old}$. Na etapa de maximização (M), as estimativas dos parâmetros são atualizadas maximizando a função de log-verossimilhança, dadas as probabilidades posteriores $\hat{P}(\zeta = c|\mathbf{Y}_i)$, tal que

$$\hat{\boldsymbol{\gamma}}_c^{new} = \frac{1}{N}\sum_{i=1}^{N}\hat{P}(\zeta = c|\mathbf{Y}_i), \tag{2.6}$$

$c = 1, \ldots, C$ definem as novas prevalências e

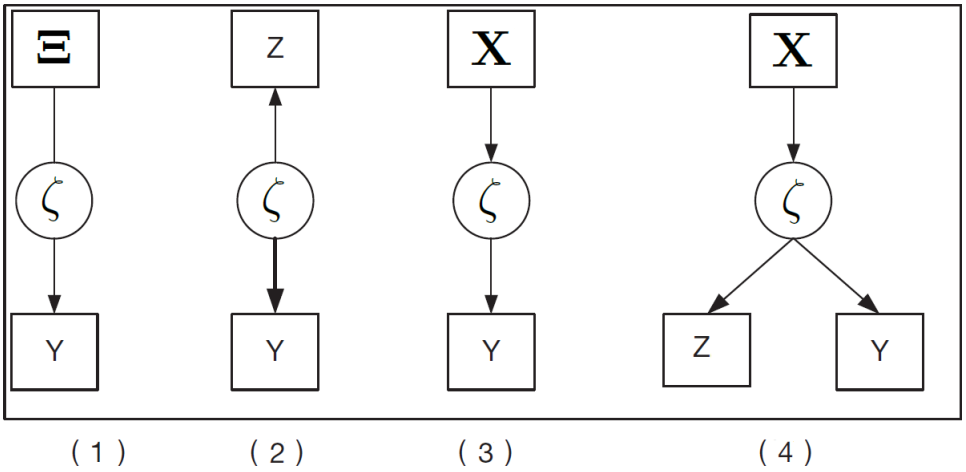
$$\hat{\boldsymbol{\rho}}_{k|c}^{new} = \frac{\sum_{i=1}^{N}\Upsilon_{ik}\hat{P}(\zeta = c|\mathbf{Y}_i))}{\sum_{i=1}^{N}\hat{P}(\zeta = c|\mathbf{Y}_i))}, \tag{2.7}$$

definem as novas probabilidades de resposta condicionais à classe latente. O vetor $\hat{\boldsymbol{\rho}}_{k|c}^{new}$ na Equação (2.7) tem comprimento R_k, contendo probabilidades de resposta condicionais à classe latente c para a k-ésima variável observada; enquanto Υ_{ik} contém os resultados observados para o k-ésimo indicador do i-ésimo indivíduo. O algoritmo repetirá estas etapas, atualizando as estimativas até que o logaritmo da função de verossimilhança seja maximizado (Linzer e Lewis, 2011).

2.2 LCA com covariáveis

Segundo Bakk, Tekle e Vermunt (2013), existem múltiplas formas pelas quais as variáveis externas podem desempenhar um papel em um modelo de LCA. De uma forma mais geral, pode-se pensar na variável de classe latente ζ sendo medida por seus indicadores Y e associada a um vetor de variáveis externas $\boldsymbol{\Xi} = (\mathbf{Z}, \mathbf{X})$, sem especificar uma ordem causal entre $\mathbf{X}$, vetor de covariáveis do modelo, e $\mathbf{Z}$, vetor de desfechos distais, como mostrado na Figura 2.1 (1). Casos mais específicos ocorrem quando: (a) Z é uma resposta distal (Figura 2.1 (2)); (b) X é preditora da variável latente ζ (Figura 2.1 (3)) ou (c) $\mathbf{X}$ é preditor de ζ, que, por sua vez, é preditor do desfecho distal Z (Figura 2.1 (4)).

Figura 2.1: Modelos de LCA com variáveis externas



Fonte: Adaptado de Bakk, Tekle e Vermunt (2013)

Para o caso em que $\mathbf{X}$ é um preditor de ζ, o modelo de LCA é também conhecido como modelo de regressão de classe latente, generalizando o modelo básico de classe latente ao permitir a inclusão de covariáveis que têm efeito sobre as prevalências de cada uma das classes. Segundo Collins e Lanza (2013), as covariáveis são incorporadas na LCA usando uma estrutura de regressão logística, e, como em qualquer estrutura de regressão, o conjunto de covariáveis pode incluir variáveis categóricas, quantitativas, ou uma combinação de ambas.

Quando covariáveis são incluídas na LCA, tem-se um modelo para $P(\mathbf{Y}_i|X_i)$ ao invés de $P(\mathbf{Y}_i)$ (Vermunt, 2010). Considerando uma covariável X, o novo modelo pode ser expresso por

$$P(\mathbf{Y}_i|X) = \sum_{c=1}^{C} \gamma_c(x) \prod_{k=1}^{K} \prod_{r_k=1}^{R_k} \rho_{k,r_k|c}^{I(Y_{ik}=r_k)}, \tag{2.8}$$

que é um modelo bastante semelhante ao definido na Equação (2.3), contudo com a presença do termo $\gamma_c(x)$, que, para o caso de uma única covariável X, pode ser definido como

$$\gamma_c(x) = P(\zeta = c|X) = \frac{e^{\vartheta_{0c}+\vartheta_{1c}x}}{1+\sum_{s=1}^{C-1} e^{\vartheta_{0s}+\vartheta_{1s}x}} \ . \tag{2.9}$$

A regressão logística multinomial requer que uma das classes da variável latente seja considerada como referência. As probabilidades de resposta ao item ainda são estimadas, mas não as prevalências das classes latentes. Neste caso, os coeficientes de regressão são estimados ($\boldsymbol{\vartheta}$) e as prevalências das classes latentes são expressas como função dos coeficientes de regressão e dos valores individuais das covariáveis correspondentes (Collins e Lanza, 2009).

2.2.1 Estimação em uma etapa

O modelo de classes latentes com vetor de variáveis externas $\mathbf{X}$ como preditoras da variável latente ζ tem o logarítmo da função de verossimilhança semelhante à Equação (2.5), exceto pela função $\gamma_c(\mathbf{x}_i)$. A nova função de log-verossimilhança pode ser expressa da seguinte forma:

$$\log(L) = \sum_{i=1}^{N} log \left[\sum_{c=1}^{C} \gamma_c(\mathbf{x}_i) \prod_{k=1}^{K} \prod_{r_k=1}^{R_k} \rho_{k,r_k|c}^{I(Y_{ik}=r_k)} \right] . \tag{2.10}$$

De acordo com Linzer e Lewis (2011), os parâmetros estimados pelo modelo de regressão de classes latentes são os $(C-1)$ vetores de coeficientes $\boldsymbol{\vartheta}_c$ e, como no modelo básico de classes latentes, as probabilidades condicionais à classe latente $\rho_{k,r_k|c}$. Dadas as estimativas $\hat{\boldsymbol{\vartheta}}_c$ e $\hat{\rho}_{k,r_k|c}$ desses parâmetros, as probabilidades posteriores de pertencimento às classes latentes são obtidas substituindo o termo γ_c da Equação (2.3) pela função $\gamma_c(\mathbf{x}_i)$. Assim:

$$\hat{P}(\zeta = c|\mathbf{X}_i; \mathbf{Y}_i) = \frac{\gamma_c(\mathbf{x}_i) \prod_{k=1}^{K} \prod_{r_k=1}^{R_k} \hat{\rho}_{k,r_k|c}^{I(Y_{ik}=r_k)}}{\sum_{c=1}^{C} \gamma_c(\mathbf{x}_i) \prod_{k=1}^{K} \prod_{r_k=1}^{R_k} \hat{\rho}_{k,r_k|c}^{I(Y_{ik}=r_k)}} , \tag{2.11}$$

com $\gamma_c(\mathbf{x}_i)$ definido em Vermunt (2010) por:

$$\gamma_c(\mathbf{x}_i) = P(\zeta = c|\mathbf{X}_i) = \frac{\exp\left(\vartheta_{0c} + \sum_{q=1}^{Q} \vartheta_{qc}\mathbf{x}_{iq}\right)}{\sum_{s=1}^{C} \exp\left(\vartheta_{0s} + \sum_{q=1}^{Q} \vartheta_{qs}\mathbf{x}_{iq}\right)} . \tag{2.12}$$

O processo de estimação começa com valores iniciais para $\hat{\boldsymbol{\vartheta}}_c^{old}$ e $\hat{\rho}_{k,r_k|c}^{old}$, que são usados para calcular as probabilidades posteriores (Linzer e Lewis, 2011). Os coeficientes de regressão são atualizados de modo que

$$\hat{\boldsymbol{\vartheta}}_c^{new} = \hat{\boldsymbol{\vartheta}}_c^{old} - H_{\vartheta}^{-1} \nabla_{\vartheta} , \tag{2.13}$$

em que ∇_{ϑ} é o vetor gradiente (vetor escore) e H_{ϑ} é a matriz hessiana da função de log-verossimilhança com respeito a ϑ. As probabilidades $\hat{\rho}_{k,r_k|c}^{new}$ são atualizadas por

$$\hat{\boldsymbol{\rho}}_{k|c}^{new} = \frac{\sum_{i=1}^{N} \boldsymbol{\Upsilon}_{ik} \hat{P}(\zeta = c|\mathbf{X}_i; \mathbf{Y}_i))}{\sum_{i=1}^{N} \hat{P}(\zeta = c|\mathbf{X}_i; \mathbf{Y}_i))} . \tag{2.14}$$

Essas etapas são repetidas até a convergência, atualizando as estimativas dos parâmetros em cada iteração. As fórmulas para vetor escore e matriz Hessiana são fornecidas em Bandeen-Roche et al. (1997).

2.2.2 Estimação em três etapas

Embora haja uma forma de estimação dos parâmetros do modelo de LCA com covariáveis em uma única etapa, nem sempre esta abordagem é utilizada na prática. Existem diferentes razões para tal escolha, sendo a mais disseminada aquela que discute as análises sendo conduzidas separadamente: (a) inicialmente pelo estabelecimento de um modelo teórico contendo os indicadores para mensurar e descrever a variável latente ζ; e (b) posteriormente, a seleção de covariáveis com objetivo de investigar sua relação com a variável latente definida em (a). Nestes casos, o método de estimação em duas etapas apresenta uma vantagem em relação à estimação simultânea dos parâmetros, pois o acréscimo de novas covariáveis em um modelo LCA requer a reestimação do modelo completo, caso seja realizada estimação simultânea (Vermunt, 2010). Adicionalmente nem sempre a escolha das covariáveis é vista pelos pesquisadores como pertencente ao mesmo estágio do estudo em que é feita a discussão do modelo teórico para a variável latente e a seleção das variáveis indicadoras. O acréscimo posterior de covariáveis no modelo pode ainda alterar a estrutura de mensuração da variável latente.

O método de estimação em três etapas é segmentado da seguinte forma: (1) estimação do modelo de mensuração utilizando informações dos padrões de resposta Y_{ik} dos indivíduos de acordo com o modelo definido na Equação (2.3); (2) alocação dos indivíduos nas classes preditas usando regras de atribuição e cálculo da quantidade de erro de classificação do modelo; e (3) estimação do modelo estrutural utilizando as classes preditas como uma variável resposta observada.

De acordo com Vermunt (2010) e Bakk, Tekle e Vermunt (2013), para atribuir ou alocar um indivíduo numa classe latente no segundo passo é necessário adotar alguma regra de classificação específica, sendo às alocações modal e proporcional as duas regras mais comuns. Na alocação modal cada indivíduo é atribuído à classe na qual possui maior probabilidade de classificação a posteriori, com a classe predita sendo denotada aqui por $\mathcal{C}$. Esta regra de alocação produz um particionamento descrito na literatura como rígido, no qual o indivíduo i é tratado como pertencente à classe c com peso de $\mathscr{c}_{ic} = P(\mathcal{C}_i = c|\mathbf{Y}_i) = 1$ se a probabilidade de classificação a posteriori $P(\zeta = c|\mathbf{Y}_i)$ for maior para a c-ésima classe em relação as demais, e com peso de $\mathscr{c}_{ic} = 0$ caso contrário. A alocação proporcional, por sua vez, produz um particionamento denominado suave, em que cada indivíduo é tratado como pertencente à classe c com probabilidade $P(\zeta = c|\mathbf{Y}_i)$, ou seja, cada indivíduo i é alocado numa classe com pesos $\mathscr{c}_{ic} = P(\mathcal{C}_i = c|\mathbf{Y}_i) = P(\zeta = c|\mathbf{Y}_i)$ (Vermunt, 2010). Caso a estimação do efeito da covariável sobre a variável latente (considerada como observada por um dos métodos de atribuição) seja feita sem correções adicionais, este procedimento é conhecido como método "*naive*" de estimação em três

etapas, pois utiliza as classes preditas como se fossem conhecidas, resultando geralmente em subestimação dos efeitos (Vermunt, 2010).

Ainda conforme Vermunt (2010), a quantidade de erro de classificação pode ser mensurada por meio da probabilidade condicional $P(\mathcal{C} = s|\zeta = c)$, probabilidade de alocar um indivíduo numa classe predita s dado que ele pertence verdadeiramente a uma classe c. O erro de classificação é definido usando regras de probabilidade condicional, de modo que

$$\begin{aligned} P(\mathcal{C} = s|\zeta = c) &= \sum_{\mathbf{Y}} P(\mathbf{Y}|\zeta = c)P(\mathcal{C} = s|\mathbf{Y}) \\ &= \frac{\sum_{\mathbf{Y}} P(\mathbf{Y})P(\zeta = c|\mathbf{Y})P(\mathcal{C} = s|\mathbf{Y})}{P(\zeta = c)} , \end{aligned} \tag{2.15}$$

com a soma ocorrendo sobre todos os padrões de resposta possíveis, e a proporção total de erros de classificação podendo ser obtida por

$$\sum_{c=1}^{C} P(\zeta = c) \sum_{s \neq c} P(\mathcal{C} = s|\zeta = c).$$

Usualmente é prático substituir a soma de todos os padrões de resposta possíveis pela soma de todos os padrões de resposta na amostra, o que implica que $P(\mathbf{Y})$ é substituído por sua distribuição empírica, o que resulta na expressão

$$\begin{aligned} P(\mathcal{C} = s|\zeta = c) &= \frac{\sum_{i=1}^{N} P(\zeta = c|\mathbf{Y}_i)P(\mathcal{C}_i = s|\mathbf{Y}_i)}{P(\zeta = c)} \\ &= \frac{\sum_{i=1}^{N} P(\zeta = c|\mathbf{Y}_i)\mathscr{c}_{is}}{P(\zeta = c)} , \end{aligned} \tag{2.16}$$

em que N denota o tamanho da amostra.

É importante ressaltar que o conceito de erro de classificação está fortemente relacionado ao conceito de separação entre as classes (Bakk, Tekle e Vermunt, 2013), que se refere a quão bem as classes podem ser distinguidas com base nas informações disponíveis sobre $\mathbf{Y}$. Mais especificamente, menor separação entre as classes corresponde a maiores erros de classificação já que torna mais incerta à alocação em uma determinada classe. Ainda conforme Bakk, Tekle e Vermunt (2013), nesse contexto, uma medida que pode ser utilizada para mensurar a separação entre as classes é baseada no princípio da entropia:

$$-\sum_{c} P(\zeta = c|\mathbf{Y} = \mathbf{y}) log P(\zeta = c|\mathbf{Y} = \mathbf{y}),$$

de modo que à medida que a entropia diminui, ocorre o mesmo com a separação entre as classes, aumentando a incerteza do procedimento de alocação dos indivíduos nas classes latentes. Contudo esta medida não é definida numa escala que permita a comparação

entre modelos, uma vez que seu valor pode variar de $[0, \infty)$. Uma alternativa a este problema é utilizar a entropia relativa (ou padronizada), que varia numa escala de $[0, 1]$. A entropia relativa é definida com mais detalhes na seção (2.4.1).

O terceiro passo envolve o ajuste de um modelo de regressão logística multinomial assumindo a classificação predita $\mathcal{C}$ como variável dependente observada/conhecida. Para sistematizar o procedimento de estimação do efeito de covariáveis na prevalência da classe predita $\mathcal{C}$, seja X_{iq} uma das Q covariáveis e $\mathbf{X}_i$ o vetor de covariáveis para um sujeito i, de modo que o seguinte modelo de regressão logística multinomial é ajustado:

$$P(\mathcal{C} = c|\mathbf{X}_i) = \frac{\exp\left(\vartheta_{0c} + \sum_{q=1}^{Q} \vartheta_{qc} X_{iq}\right)}{\sum_{s=1}^{C} \exp\left(\vartheta_{0s} + \sum_{q=1}^{Q} \vartheta_{qs} X_{iq}\right)} , \tag{2.17}$$

em que os parâmetros de interesse são os ϑ_{qc}, para $0 \leq q \leq Q$, que são estimados através da maximização da seguinte função de log-verossimilhança ponderada:

$$\log(L_{STEP3}) = \sum_{i=1}^{N} \sum_{c=1}^{C} \mathscr{c}_{ic} \log(P(\mathcal{C} = c|\mathbf{X}_i)) , \tag{2.18}$$

com $\mathscr{c}_{ic} = P(\mathcal{C} = c|\mathbf{Y}_i)$. Este processo de estimação equivale ao ajuste de uma regressão logística multinomial sobre um conjunto de dados expandido com C registros por unidade amostral e com $\mathscr{c}_{ic}$ como pesos (Vermunt, 2010).

Como já mencionado, é comum que muitos pesquisadores optem pela metodologia de estimação em três etapas, estabelecendo inicialmente um modelo de mensuração para posteriormente investigarem as relações entre as covariáveis e variáveis latentes. Assim, diversos autores (Bakk e Kuha, 2021; Bakk, Oberski e Vermunt, 2016; Bakk, Tekle e Vermunt, 2013; Bolck, Croon e Hagenaars, 2004; Bray, Lanza e Tan, 2015; Lanza, Tan e Bray, 2013; Vermunt, 2010) têm realizado estudos de simulação para avaliar o viés dos estimadores que pode ser introduzido por ignorar o erro de classificação quando se utiliza as classes preditas $\mathcal{C}$ ao invés das classes latentes verdadeiras ζ.

No âmbito deste trabalho, o foco está voltado para o modelo de LCA com desfechos distais. Deste modo, a discussão sobre os métodos de correção de viés na estimação em três etapas, que visam incorporar os erros de classificação, são apresentados na subseção 2.3.

2.3 LCA com desfecho distal

Assim como foi comentado na Seção (2.2), é possível realizar a incorporação de variáveis externas $\boldsymbol{\Xi} = (\mathbf{Z}, \mathbf{X})$ no modelo de LCA, e para o caso em que a ordem causal é definida como nas Figuras 2.1(2) e 2.1(4), com a variável latente ζ sendo preditora do

desfecho distal Z. Desta forma, tem-se o caso de um modelo de LCA com desfecho distal. De acordo com Bakk, Tekle e Vermunt (2013), a extensão deste tipo de modelo pode ser composto por duas partes: o modelo de mensuração, que incorpora informações acerca da variável latente ζ e seus indicadores $\mathbf{Y}$, e o modelo estrutural que incorpora as relações entre ζ e Z.

Em Bakk e Kuha (2021) é apresentada a formulação teórica do modelo de LCA com variáveis externas, incorporando a informação tanto dos preditores observados (covariáveis) para ζ, denotados por $\mathbf{X}_i$ quanto do desfecho distal denotado por Z_i, culminando na seguinte equação:

$$P(\mathbf{Y}_i, Z_i|\mathbf{X}_i) = \sum_{c=1}^{C} \left[P(Z_i, \zeta = c|\mathbf{X}_i) \prod_{k=1}^{K} P(Y_{ik}|\zeta = c) \right] , \tag{2.19}$$

em que

$$P(Z_i, \zeta = c|\mathbf{X}_i) = P(\zeta = c|\mathbf{X}_i) P(Z_i|\zeta = c, \mathbf{X}_i) . \tag{2.20}$$

Em muitas aplicações, no entanto, o modelo de interesse incluirá apenas covariáveis no modelo de classes latentes através do termo $P(\zeta = c|\mathbf{X}_i)$ (Equação (2.20)) ou apenas as respostas distais através do termo $P(Z_i|\zeta = c, \mathbf{X}_i)$ (Equação (2.20)). Há ainda a possibilidade de extensão do modelo definido na Equação (2.19), de modo que o modelo estrutural englobe múltiplas variáveis latentes, múltiplas respostas distais e/ou algumas variáveis que são covariáveis em alguns modelos e respostas distais em outros (Bakk e Kuha, 2021). Neste trabalho considera-se um modelo contendo apenas uma variável latente e uma resposta distal.

2.3.1 Estimação em uma etapa em LCA com desfecho distal

Segundo Bakk, Tekle e Vermunt (2013), a forma mais geral de associação entre variáveis externas e a variável latente ζ sem especificar uma ordem causal como na Figura 2.1 (1) envolve modelar a seguinte probabilidade conjunta:

$$P(\Xi, \zeta = c, \mathbf{Y}_i) = P(\Xi, \zeta = c) P(\mathbf{Y}_i|\zeta = c) . \tag{2.21}$$

Na Equação (2.21) assume-se que a variável externa Ξ e Y são condicionalmente independentes dado ζ, sendo possível sua adaptação para casos específicos. Quando ζ é preditora da variável externa, como na Figura 2.1 (2), o modelo pode ser definido com a seguinte formulação (Bakk, Tekle e Vermunt, 2013):

$$P(Z_i, \zeta = c, \mathbf{Y}_i) = P(\zeta = c)P(Z_i|\zeta = c)P(\mathbf{Y}_i|\zeta = c) \ . \tag{2.22}$$

Note que, ao usar uma variável externa como desfecho distal, utiliza-se a notação Z ao invés de Ξ. Para um caso mais geral em que haja covariáveis e desfechos distais, o novo modelo pode ser definido da seguinte forma:

$$P(\mathbf{Y}_i, Z_i, \zeta = c|\mathbf{X}_i) = P(\zeta = c|\mathbf{X}_i)P(Z_i|\zeta = c, \mathbf{X}_i)P(\mathbf{Y}_i|\zeta = c) \ . \tag{2.23}$$

Esta formulação é análoga à definição completa para este tipo de modelo, apresentada em Bakk e Kuha (2021), pois ao realizar uma soma sobre todas as C classes latentes, tem-se:

$$\begin{aligned} P(\mathbf{Y}_i, Z_i|\mathbf{X}_i) &= \sum_{c=1}^{C} P(\mathbf{Y}_i, Z_i, \zeta = c|\mathbf{X}_i) \\ &= \sum_{c=1}^{C} [P(\zeta = c|\mathbf{X}_i)P(Z_i|\zeta = c, \mathbf{X}_i)P(\mathbf{Y}_i|\zeta = c)] \\ &= \sum_{c=1}^{C} [P(Z_i, \zeta = c|\mathbf{X}_i)P(\mathbf{Y}_i|\zeta = c)] \\ &= \sum_{c=1}^{C} \left[P(Z_i, \zeta = c|\mathbf{X}_i) \prod_{k=1}^{K} P(Y_{ik}|\zeta = c) \right] \ . \end{aligned} \tag{2.24}$$

Assim, é possível escrever a função de log-verossimilhança, como apresentado nas Seções (2.1.1) e (2.2.1), da seguinte forma:

$$\log(L) = \sum_{i=1}^{N} log \left[\sum_{c=1}^{C} \left(P(Z_i, \zeta = c|\mathbf{X}_i) \prod_{k=1}^{K} P(Y_{ik}|\zeta = c) \right) \right] \ , \tag{2.25}$$

com os parâmetros do modelo sendo obtidos através da maximização do logaritmo da verossimilhança acima, em que $P(Z_i, \zeta = c|\mathbf{X}_i)$ pode ser estimado empiricamente (Bakk e Kuha, 2021).

2.3.1.1 Método de correção LTB

Como alternativa à estimação em uma etapa, para o caso de um modelo com desfecho distal foi desenvolvido o método de correção LTB. Este método foi inicialmente apresentado em Lanza, Tan e Bray (2013), e denominado como método de correção de viés LTB, de acordo com as iniciais dos seus autores. Embora este método tenha sido criado como contraponto às abordagens de estimação em três etapas discutidas nas Seções (2.2.2)

e (2.3.2), ele foi idealizado para ser estimado em uma etapa, sem que fosse necessário utilizar algum método de alocação, evitando assim o viés causado pela utilização de $\mathcal{C}$ ao invés de ζ.

Segundo Lanza, Tan e Bray (2013), embora haja um interesse considerável em estimar a relação entre uma variável de classe latente, ζ, e uma resposta distal, Z, é preciso adotar adicionalmente a suposição de independência condicional entre $\mathbf{Y}$ e Z dado ζ para que esta estimação seja possível. Assim, considera-se que $P(\mathbf{Y}, Z|\zeta) = P(\mathbf{Y}|\zeta)P(Z|\zeta)$.

De acordo com Bakk e Kuha (2021), a ideia por trás deste tipo de estimação é tratar o desfecho distal Z como uma covariável de ζ num modelo de LCA, culminando na seguinte equação:

$$\begin{aligned} P(\mathbf{Y}_i, Z) &= \sum_{c=1}^{C} \left[P(Z, \zeta = c) \prod_{k=1}^{K} P(Y_{ik}|\zeta = c) \right] \\ &= P(Z) \sum_{c=1}^{C} \left[P(\zeta = c|Z) \prod_{k=1}^{K} P(Y_{ik}|\zeta = c) \right] , \end{aligned} \tag{2.26}$$

com $P(Z)$ denotando a distribuição marginal de Z. Ainda segundo Bakk e Kuha (2021), este método tem a vantagem de não necessitar de suposições distributivas sobre Z, cuja especificação incorreta pode influenciar as estimativas do modelo. A distribuição condicional de ζ dado Z pode ser especificada como um modelo logístico multinomial, como apresentado na Seção (2.2.2).

Para que seja possível obter $P(Z|\zeta)$, que é a distribuição de interesse, Lanza, Tan e Bray (2013) propuseram utilizar a seguinte relação baseada no teorema de Bayes:

$$P(Z|\zeta = c) = \frac{P(Z)P(\zeta = c|Z)}{P(\zeta = c)} .$$

Embora Lanza, Tan e Bray (2013) não tenham proposto estimadores para os erros padrão, outros autores, como Asparouhov e Muthén (2014a) e Asparouhov e Muthén (2014b), sugeriram usar o método delta para respostas distais categóricas e erros padrão aproximados para respostas distais contínuas.

2.3.2 Estimação em três etapas em LCA com desfecho distal

Como já descrito na Seção (2.2.2), na abordagem conhecida como estimação em três etapas, inicialmente os parâmetros do modelo de mensuração são estimados a partir de indicadores de classe latente pré-estabelecidos. Em um segundo passo, os indivíduos são alocados em classes latentes de acordo com seus vetores de padrão de resposta $\mathbf{Y}_i$. E, por fim, são estimados os parâmetros do modelo estrutural. De acordo com Bakk e Kuha (2021), no terceiro passo as classes em que os indivíduos são alocados, representadas por

$\mathcal{C}$, são utilizadas no papel de ζ para estimação do modelo estrutural. No método clássico (*naive*), isso é feito sem nenhum ajuste adicional. Por exemplo, os componentes do modelo com covariáveis e desfecho distal são simplesmente substituídos por:

$$P(Z_i, \mathcal{C}|\mathbf{X}_i) = P(\mathcal{C}|\mathbf{X}_i)P(Z_i|\mathcal{C}, \mathbf{X}_i).$$

Desta forma os termos $P(\mathcal{C}|\mathbf{X}_i)$ e $P(Z_i|\mathcal{C}, \mathbf{X}_i)$ são análogos a $P(\zeta = c|\mathbf{X}_i)$ e $P(Z_i|\zeta = c, \mathbf{X}_i)$ respectivamente, e geralmente são modelados através de uma regressão logística multinomial para relação entre $\mathcal{C}$ e $\mathbf{X}_i$, e uma regressão linear considerando um Z contínuo. O problema com esta abordagem, conhecida como "naive", é que $\mathcal{C}$ não é necessariamente igual a ζ. A atribuição na etapa 2, portanto, introduz um erro de classificação que pode enviesar severamente as estimativas da etapa 3, como foi discutido na Seção 2.2.2.

A chave para os métodos de correção reside no fato de que é possível mostrar como a distribuição (ζ, Ξ) está relacionada à distribuição $(\mathcal{C}, \Xi)$. Para o caso em que $\Xi = (Z, \mathbf{X})$, ou seja, desfecho distal e covariáveis, as seguintes igualdades foram desenvolvidas, respectivamente, por Bakk, Tekle e Vermunt (2013) e Bakk e Kuha (2021):

$$P(Z = z, \mathcal{C} = s) = \sum_c P(Z = z, \zeta = c)P(\mathcal{C} = s|\zeta = c), \tag{2.27}$$

$$P(Z = z, \mathcal{C} = s|\mathbf{X}_i) = \sum_c P(Z = z, \zeta = c|\mathbf{X}_i)P(\mathcal{C} = s|\zeta = c). \tag{2.28}$$

A partir da distribuição conjunta apresentada na Equação (2.27), a distribuição condicional de Z dado ζ pode ser obtida quando a variável latente ζ é considerada um preditor da variável externa Z (Bakk, Tekle e Vermunt, 2013). Já na Equação (2.28) a distribuição empírica $P(Z = z, \mathcal{C} = s|\mathbf{X}_i)$ é utilizada para estimar $P(Z = z, \zeta = c|\mathbf{X}_i)$, com as $P(\mathcal{C} = s|\zeta = c)$ sendo conhecidas a partir das etapas 1 e 2 da estimação (Bakk e Kuha, 2021). Desta forma a extensão dos métodos reside na constatação de que o erro de classificação depende apenas do modelo de mensuração.

Os primeiros métodos foram desenvolvidos ainda para o caso do modelo de LCA com covariáveis, como o método de correção de viés desenvolvido por Bolck, Croon e Hagenaars (BCH) (Bolck, Croon e Hagenaars, 2004) e modificado por Vermunt (2010), e o método de correção de viés baseado em máxima verossimilhança (ML) (Vermunt, 2010). Estes métodos trouxeram contribuições importantes, mostrando que era possível diminuir o viés das estimativas do modelo quando comparados com o método clássico (*naive*) de estimação em três etapas. Embora estes métodos tenham sido inicialmente propostos para o caso em que havia um modelo de LCA com covariáveis, Bakk e Kuha (2021) demonstraram que estes métodos podem ser também adaptados para o caso de um modelo contendo desfechos distais, diminuindo assim o viés das estimativas. Outros

autores propuseram também métodos específicos para modelos com desfechos distais, como é o caso do método LTB proposto inicialmente em Lanza, Tan e Bray (2013) e adaptado por Bakk, Oberski e Vermunt (2016) para ser utilizado numa abordagem de três etapas.

2.3.2.1 Método de correção BCH

Este método de correção foi desenvolvido para o cenário em que haviam variáveis categóricas externas como preditoras, ou seja, em modelos com preditores $\mathbf{X}_i$ para ζ. Este método envolve reexpressar a probabilidade conjunta entre a variável latente ζ e um vetor de preditores $\mathbf{X}_i$ como:

$$P(\zeta = c, \mathbf{X}_i) = \sum_s P(\mathcal{C} = s, \mathbf{X}_i) d^*_{sc} \ ,$$

em que d^*_{sc} representa os elementos da matriz inversa $\mathbf{D}^{-1}$, com $\mathbf{D}$ sendo definida como uma matriz de dimensão $C \times C$ com elementos $P(\mathcal{C} = s, \zeta = c)$. A ideia é ponderar a distribuição conjunta $(\mathcal{C}, Z)$ pelo inverso dos erros de classificação para obter a distribuição de interesse. É importante destacar que para que essa relação seja possível, a matriz $\mathbf{D}$ deve ser não singular, exigindo a condição de que $P(\mathcal{C} = s|\zeta = c) = P(\mathcal{C} = s|\zeta = c'), \forall s$, não seja válida para qualquer $c \neq c'$ (Vermunt, 2010).

Para este método de correção em três etapas é importante que todas as covariáveis sejam categóricas, o que implica que os dados podem ser resumidos em uma tabela de contingência. Desta forma, para $\mathbf{X}^*_j$ denotando um dos j padrões de covariáveis, e n_{js} o número de indivíduos ou observações com padrão de resposta j que foram alocados na classe s, Bakk, Tekle e Vermunt (2013) mostram que esta abordagem envolve a maximização da seguinte função de pseudo log verossimilhança:

$$\begin{aligned} \log(L_{BCH}) &= \sum_j \sum_s n_{js} \sum_c d^*_{sc} \log(P(\zeta = c, X = x)) \\ &= \sum_j \sum_c n^*_{jc} \log(P(\zeta = c, X = x)) \ , \end{aligned}$$

com $n^*_{jc} = \sum_{s=1}^{C} n_{js} d^*_{sc}$ representando as frequências reponderadas usadas para estimar a relação entre ζ e $\mathbf{X}$ (Bolck, Croon e Hagenaars, 2004). Segundo Bakk, Tekle e Vermunt (2013), esta relação, que se aplica ao nível da população, é usada para reponderar os dados em $\mathcal{C}$ e $\mathbf{X}$.

2.3.2.1.1 Método de correção BCH modificado

De acordo com Vermunt (2010), o método de correção BCH tem três limitações principais: só pode ser usado com preditores $\mathbf{X}$ categóricos, uma nova matriz de dados deve ser criada cada vez que o conjunto de covariáveis sob investigação é alterado, e o procedimento não produz erros-padrão corretos. Para resolver estas limitações foi proposta uma nova abordagem que consiste em reexpressar a função de pseudo log-verossimilhança em termos de observações individuais, da seguinte forma:

$$\begin{aligned}\log(L_{BCH}) &= \sum_{i=1}^{N}\sum_{s=1}^{C} \ddot{c}_{is} \sum_{c}^{C} d^*_{sc} \log(P(\zeta = c|\mathbf{X}_i)) \\ &= \sum_{i=1}^{N}\sum_{c}^{C} \ddot{c}_{ic} \log(P(\zeta = c|\mathbf{X}_i)) \ ,\end{aligned}$$

com $\ddot{c}_{is}$ sendo o peso de alocação em uma classe s, de acordo com a regra de alocação escolhida e $\ddot{c}_{ic} = \sum_s \ddot{c}_{is} d^*_{sc}$.

Contudo, para que fosse possível realizar a extensão deste método de correção para o caso de haver um desfecho distal Z, Bakk, Tekle e Vermunt (2013), diferentemente de Vermunt (2010), reexpressaram a função de log-verossimilhança com base da distribuição conjunta de $\mathcal{C}$ e uma variável externa, aqui denotada como Z por se tratar de um desfecho distal, resultando na Equação (2.29):

$$\begin{aligned}\log(L_{BCH}) &= \sum_{i=1}^{N}\sum_{s=1}^{C} \ddot{c}_{is} \sum_{c}^{C} d^*_{sc} \log(P(\zeta = c, Z = z_i)) \\ &= \sum_{i=1}^{N}\sum_{c}^{C} \ddot{c}_{ic} \log(P(\zeta = c, Z = z_i)) \ ,\end{aligned} \tag{2.29}$$

e então esta equação pode ser modificada para a seguinte forma:

$$\begin{aligned}\log(L_{BCH}) &= \sum_{i=1}^{N}\sum_{c}^{C} \ddot{c}_{ic} \log(P(\zeta = c)P(Z = z|\zeta = c)) \\ &= \sum_{i=1}^{N}\sum_{c}^{C} \ddot{c}_{ic} \log(P(\zeta = c)) + \sum_{i=1}^{N}\sum_{c}^{C} \ddot{c}_{ic} \log(P(Z = z|\zeta = c)) \ .\end{aligned} \tag{2.30}$$

Segundo Bakk, Tekle e Vermunt (2013), como o primeiro termo da Equação (2.30) não contém parâmetros de interesse, ele pode ser ignorado, sendo maximizada apenas uma função de pseudo log-verossimilhança com base no segundo termo. O autor também destaca que esta formulação permite a aplicação do método BCH modificado a variáveis

externas de qualquer tipo de escala e, portanto, também a variáveis Z contínuas e ordinais. Bakk, Tekle e Vermunt (2013) ainda afirmam que, ao aplicar um estimador de variância robusto ou sanduíche, se pode evitar que os erros-padrão sejam subestimados, como é o caso da abordagem BCH original. A matriz de variância-covariância robusta dos parâmetros é definida pelo inverso da matriz obtida pelo estimador "sanduíche" usando a matriz Hessiana e a média do produto externo dos gradientes para as observações independentes (Bakk, Tekle e Vermunt, 2013; Skinner, Holt e Smith, 1989).

2.3.2.2 Método de correção ML

De acordo com Bakk, Tekle e Vermunt (2013), o método de correção baseado em máxima verossimilhança (ML, em inglês), que foi apresentado inicialmente em Vermunt (2010), envolve a definição de um modelo de classe latente com uma ou mais covariáveis (X) afetando a variável latente ζ e tendo $\mathcal{C}$ como o único indicador da variável latente desconhecida ζ. Uma diferença importante em comparação com um LCA clássico é que $P(\mathcal{C} = s|\zeta = c)$ são fixadas em seus valores estimados da etapa anterior. Ainda segundo Bakk, Tekle e Vermunt (2013), o método de correção ML pode ser facilmente adaptado para a modelagem da distribuição conjunta de ζ e Z ou da distribuição condicional de Z dado ζ. Desta forma a Equação (2.27) também pode ser reescrita da seguinte forma:

$$P(Z = z, \mathcal{C} = s) = \sum_c P(\zeta = c)P(Z = z|\zeta = c)P(\mathcal{C} = s|\zeta = c) \tag{2.31}$$

para a situação em que ζ é um preditor de Z, que é o desfecho distal. Uma suposição subjacente a esse modelo é que Z e $\mathcal{C}$ são condicionalmente independentes dado ζ, suposição que é necessária para todas as abordagens de três etapas atualmente existentes. É ainda necessário que se especifique a forma da distribuição de $P(Z = z|\zeta = c)$.

Assim os parâmetros do modelo apresentado na Equação (2.31) podem ser estimados maximizando a equação:

$$\log(L_{ML}) = \sum_i \log \sum_c P(\zeta = c)P(Z = z|\zeta = c)P(\mathcal{C} = s|\zeta = c).$$

2.4 Inferência bayesiana em análise de classes latentes

Inferência Bayesiana é uma abordagem utilizada para estimação de parâmetros de um modelo estatístico levando em consideração tanto a informação dos dados observados

quanto as informações prévias (ou a priori) que temos sobre esses parâmetros. Nesta abordagem, o objetivo é obter a distribuição a *posteriori* dos parâmetros do modelo, condicionais aos dados observados e a informações prévias disponíveis.

Segundo O'Hagan (1994), a distribuição a *posteriori* $P(\tau|\mathbf{\Omega})$ para um parâmetro de interesse τ incorpora tudo o que é conhecido sobre o parâmetro de acordo com o conjunto dos dados observados $\mathbf{\Omega}$, representados pela verossimilhança $P(\mathbf{\Omega}|\tau)$, e a informação incorporada através da *priori* $P(\tau)$. Desta forma, todos os graus posteriores de incerteza relativos ao parâmetro são expressos nela. A partir da utilização do teorema de Bayes é possível expressar a relação entre a distribuição *a posteriori*, a verossimilhança e a distribuição *a priori* da seguinte forma:

$$P(\tau|\mathbf{\Omega}) \propto P(\mathbf{\Omega}|\tau)P(\tau),$$

em que $\propto$ representa proporcionalidade. A inferência bayesiana tem se tornado cada vez mais atrativa, principalmente devido aos avanços em métodos computacionais incluindo, por exemplo, os métodos de Monte Carlo via Cadeias de Markov (MCMC) (Gilks, Richardson e Spiegelhalter, 1996) e Monte Carlo Hamiltoniano (HMC) (Duane et al., 1987), além da disponibilidade de softwares especializados, como o WinBUGS (Spiegelhalter et al., 2002), JAGS (Plummer, 2017) e STAN (Gelman, Lee e Guo, 2015; Stan Development Team, 2024b).

Na literatura sobre modelagem com variáveis latentes, as vantagens da abordagem bayesiana em relação à frequentista para o caso de modelos de equações estruturais (SEM, em inglês) têm sido discutidas, ressaltando-se a obtenção de estimativas mais confiáveis para amostras pequenas (Lee, 2007; Palomo, Dunson e Bollen, 2007). No contexto de análise de classes latentes (LCA), a abordagem bayesiana tem se tornado cada vez mais popular, especialmente pela incorporação da incerteza no processo de estimação dos parâmetros por meio de distribuições de probabilidade *a priori* (Costa, Amorim e Bispo, 2021).

Asparouhov e Muthén (2011) descreveram, dentre as vantagens desta abordagem no processo de estimação na LCA, a facilidade e a flexibilidade de acomodar especificações incorretas do modelo causadas pela violação do pressuposto de independência local, que ocorre quando dentro de cada classe as variáveis indicadoras observadas não são independentes umas das outras. Assim, esta abordagem pode evitar a formação de classes espúrias causadas por violações deste pressuposto ao relaxá-lo para uma suposição de independência aproximada. Além disso, a abordagem bayesiana no contexto de LCA facilita a obtenção de estimativas intervalares para os parâmetros do modelo de mensuração, fornecendo informações mais detalhadas sobre a incerteza associada a esses parâmetros em comparação com a abordagem frequentista.

Recentemente diversos autores têm proposto extensões e apresentado aplicações da abordagem bayesiana em modelos de classes latentes (Asparouhov e Muthén, 2011; Costa, Amorim e Bispo, 2021; Li et al., 2018; A. White e Murphy, 2014). Como exemplo de aplicações da abordagem em dados reais, Asparouhov e Muthén (2011) ilustram seu uso na análise de dados da área de saúde, enquanto Costa, Amorim e Bispo (2021) utilizam dados educacionais.

2.4.1 Parametrização para a abordagem bayesiana

De acordo com Li et al. (2018) e Costa, Amorim e Bispo (2021), o modelo clássico de classes latentes pode ser reescrito em termos da função de log-verossimilhança para os dados completos. Seja $\boldsymbol{\zeta}_i = (\zeta_{i1}, \ldots, \zeta_{iC})'$ o vetor de respostas multinomiais relacionado à variável latente para o i-ésimo indivíduo, com valores observados $\boldsymbol{\varsigma}_i = (\varsigma_{i1}, \ldots, \varsigma_{iC})'$, de forma que $\zeta_{ic} = 1$ se o indivíduo i pertence a classe c e igual a 0 caso contrário, e $P(\zeta_{ic} = 1) = \gamma_c$ para $c = 1, \cdots, C$, com $\boldsymbol{\zeta} = (\boldsymbol{\zeta}_1, \cdots, \boldsymbol{\zeta}_N)'$ (Costa, Amorim e Bispo, 2021). Considerando a função de verossimilhança dos dados completos para o i-ésimo indivíduo, o modelo de LCA clássico pode ser também definido da seguinte forma:

$$P(\mathbf{Y}_i = \mathbf{y}_i, \boldsymbol{\zeta}_i = \boldsymbol{\varsigma}_i | \boldsymbol{\gamma}, \boldsymbol{\rho}) = \prod_{c=1}^{C} \left[\gamma_c \prod_{k=1}^{K} \prod_{r_k=1}^{R_k} \rho_{k,r_k|c}^{I(y_{ik}=r_k)} \right]^{\varsigma_{ic}}, \tag{2.32}$$

com a contribuição de cada indivíduo sendo computada de acordo com o seu pertencimento à classe c.

Ao utilizar o teorema de Bayes, é possível escrever a distribuição a *posteriori* conjunta dos parâmetros $\boldsymbol{\gamma}$ e $\boldsymbol{\rho}$, condicional aos indicadores $\mathbf{Y}$ e à variável latente $\boldsymbol{\zeta}$, por meio da seguinte relação:

$$P(\boldsymbol{\gamma}, \boldsymbol{\rho} | \mathbf{Y}, \boldsymbol{\zeta}) = \frac{P(\mathbf{Y}, \boldsymbol{\zeta} | \boldsymbol{\gamma}, \boldsymbol{\rho}) P(\boldsymbol{\gamma}, \boldsymbol{\rho})}{P(\mathbf{Y}, \boldsymbol{\zeta})},$$

em que $P(\mathbf{Y}, \boldsymbol{\zeta} | \boldsymbol{\gamma}, \boldsymbol{\rho})$ representa a função da verossimilhança descrita na Equação (2.32), $P(\boldsymbol{\gamma}, \boldsymbol{\rho})$ representa *a priori* conjunta para $\boldsymbol{\gamma}$ e $\boldsymbol{\rho}$, e $P(\mathbf{Y}, \boldsymbol{\zeta})$ é a probabilidade preditiva de observar um padrão de resposta $\mathbf{y}_i$ em uma classe latente c. Ao assumir independência a priori entre os parâmetros $\boldsymbol{\gamma}$ e $\boldsymbol{\rho}$, e desconsiderando o termo $P(\mathbf{Y}, \boldsymbol{\zeta})$, também referido como constante normalizadora uma vez que não depende dos parâmetros, a distribuição *a posteriori* conjunta pode ser reescrita em termos da proporcionalidade:

$$P(\boldsymbol{\gamma}, \boldsymbol{\rho} | \mathbf{Y}, \boldsymbol{\zeta}) \propto P(\mathbf{Y}, \boldsymbol{\zeta} | \boldsymbol{\gamma}, \boldsymbol{\rho}) P(\boldsymbol{\gamma}) P(\boldsymbol{\rho}).$$

De acordo com a nova parametrização do modelo, é coerente assumir que os vetores de parâmeros $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_C)'$ e $\boldsymbol{\rho}_{k,c} = (\rho_{k,1|c}, \cdots, \rho_{k,R_k|c})'$ seguem ambos distribuições *a priori* de Dirichlet, com vetores de hiperparâmetros $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_C)'$ e $\boldsymbol{u}_{k,c} = (u_{k,1|c}, \cdots, u_{k,R_k|c})$, respectivamente (Costa, Amorim e Bispo, 2021). Os hiperparâmetros recebem esta denominação por serem parâmetros estimáveis, que compõem as distribuições de probabilidade *a priori*. Desta forma, a função de densidade de probabilidade conjunta *a priori* para $\boldsymbol{\rho} = (\boldsymbol{\rho}_1, \cdots, \boldsymbol{\rho}_C)'$, em que $\boldsymbol{\rho}_C = (\boldsymbol{\rho}_{1,1}, \cdots, \boldsymbol{\rho}_{K,C})'$, é definida como

$$P(\boldsymbol{\rho}) \propto \prod_{c=1}^{C} \prod_{k=1}^{K} \prod_{r_k=1}^{R_k} \rho_{k,r_k|c}^{u_{k,r_k|c}-1} \ ,$$

enquanto que a função de densidade de probabilidade conjunta *a priori* para γ é dada por:

$$P(\boldsymbol{\gamma}) \propto \prod_{c=1}^{C} \gamma_c^{\alpha_c - 1} \ .$$

A distribuição de Dirichlet é escolhida como distribuição *a priori* para os vetores $\boldsymbol{\gamma}$ e $\boldsymbol{\rho}_{k,c}$, com hiperparâmetros $\boldsymbol{\alpha}$ e $\boldsymbol{u}_{k,c}$, respectivamente. Essa escolha é motivada por suas propriedades conjugadas em relação à distribuição Multinomial, o que facilita a inferência bayesiana. Como $\boldsymbol{\gamma}$ representa as proporções associadas às classes latentes, cuja soma é restrita a 1, isso implica na distribuição Dirichlet como uma escolha natural para descrever incertezas sobre essas proporções. Analogamente, para $\boldsymbol{\rho}_{k,c}$, a Dirichlet é apropriada porque modela as probabilidades condicionais de cada categoria dos indicadores, que também são restritas a somar 1 dentro de cada classe. Os parâmetros $\boldsymbol{\alpha}$ e $\boldsymbol{u}_{k,c}$ permitem incorporar informações prévias, como equilíbrio ou preferências entre classes e categorias, caso estejam disponíveis.

Além disto, utilizando parametrização para o modelo de LCA, conforme Equação (2.32), é possível obter a probabilidade a posteriori de um indivíduo i pertencer a uma determinada classe latente, condicional ao seu padrão de resposta e aos parâmetros do modelo, por meio da Equação (2.33):

$$P(\boldsymbol{\zeta}_i = \boldsymbol{\varsigma}_i | \mathbf{Y}_i, \boldsymbol{\gamma}, \boldsymbol{\rho}) = \frac{\prod_{c=1}^{C} \left[\gamma_c \prod_{k=1}^{K} \prod_{r_k=1}^{R_k} \rho_{k,r_k|c}^{I(y_{ik}=r_k)}\right]^{\varsigma_{ic}}}{\sum_{c=1}^{C} \gamma_c \prod_{k=1}^{K} \prod_{r_k=1}^{R_k} \rho_{k,r_k|c}^{I(y_{ik}=r_k)}} \tag{2.33}$$

Inerente à matriz de probabilidades posteriores, que apresenta as probabilidades dos indivíduos pertencerem a cada uma das c classes, está associado o conceito de entropia do modelo na LCA. A entropia, denotada por E, é fundamental para avaliar a qualidade da classificação de um modelo de classes latentes. Quando as probabilidades

de pertencimento a uma classe são altas em relação às demais, melhor é a separação das classes e menor o grau de incerteza na classificação. Por outro lado, probabilidades semelhantes entre as classes sugerem baixa separação e homogeneidade, indicando que diferentes classes possuem perfis de resposta semelhantes (Collins e Lanza, 2009). Uma baixa separação e homogeneidade de classes também está relacionada com o aumento do erro de classificação, como discutido na Seção (2.2.2). Nesse contexto, um bom modelo idealmente apresentará alta homogeneidade e separação, refletida em probabilidades de resposta ao item próximas a 0 ou 1 condicionais à uma classe latente c.

Assim, a capacidade do modelo em classificar indivíduos está diretamente relacionada à sua separação e homogeneidade, sendo essencial avaliar as probabilidades a posteriori para cada indivíduo. Desta forma, o grau de incerteza na classificação de um indivíduo pode ser calculado através da definição de entropia relativa ou padronizada, definida inicialmente por Celeux e Soromenho (1996), e que pode ser calculada no contexto de LCA da seguinte forma:

$$E = 1 - \frac{\sum_{i=1}^{N} \sum_{c=1}^{C} -p_{ic} \log(p_{ic})}{N \log(C)} \tag{2.34}$$

em que p_{ic} é a probabilidade a posteriori do indivíduo i pertencer à c-ésima classe latente, calculada de acordo com a Equação (2.33). O valor para E varia de 0 a 1, sendo que maiores valores indicam uma melhor separação das classes latentes. De acordo com Collins e Lanza (2009), o erro de classificação também pode aumentar de acordo com o número de classes latentes consideradas, pois à medida em que este número aumenta, a estimativa para E tende a diminuir. Porém, mesmo que o valor para entropia esteja próximo a 1, ainda é possível que haja alguns indivíduos classificados erroneamente. Este tipo de entropia é uma medida útil para avaliação de um modelo de classes latentes. Além disto esta definição para E também é implementada via *software* Mplus (L. K. Muthén e Muthén, 2017), que é um consolidado programa estatístico desenvolvido especificamente para modelagem com variáveis latentes. Contudo, além da entropia padronizada é comum na prática realizar o cálculo de outras estatísticas relacionadas ao modelo de mensuração.

Usualmente há interesse em investigar qual o melhor modelo variando o número de classes latentes C. Neste caso, um modelo pode ser escolhido através do Critério de Informação de Desvio (DIC, em inglês *Deviance Information Criterion*) (Spiegelhalter et al., 2002), Critério de Informação de Akaike Monte Carlo (AICM, em inglês *Akaike Information Criterion Monte Carlo*) (Akaike, 1998) e Critério de Informação Bayesiano Monte Carlo (BICM, em inglês *Bayesian Information Criterion Monte Carlo*) (Schwarz, 1978), sendo o melhor modelo aquele que apresentar o menor valor nos critérios de informação mencionados, e o maior valor para entropia E.

2.4.2 Estimação bayesiana em análise de classes latentes

Na inferência Bayesiana, atualizar uma distribuição *a priori* por meio de uma distribuição a *posteriori* pode envolver cálculos complexos, como integrais, que podem ser analiticamente intratáveis em várias circunstâncias (Costa, Amorim e Bispo, 2021). Métodos numéricos, como os métodos MCMC e HMC, são úteis para resolver esses problemas, gerando amostras da distribuição a *posteriori* (Barbosa et al., 2010; Costa, Amorim e Bispo, 2021; Gilks, Richardson e Spiegelhalter, 1996). O método MCMC aproxima a integral que atualiza a distribuição *a priori* a partir de uma distribuição *a posteriori*. Nesse processo, a cadeia de Markov gera uma sequência de variáveis aleatórias, onde os valores amostrados na iteração $w+1$ dependem apenas dos da iteração w (Bispo, 2019). A distribuição limite dessa cadeia corresponde à distribuição *a posteriori* conjunta dos parâmetros, permitindo a inferência sobre eles (Smith e Roberts, 1993).

2.4.2.1 Amostrador de Gibbs

Neste contexto, a sigla MCMC refere-se a um conjunto de métodos utilizados para geração de amostras aleatórias, sendo o algoritmo de Metropolis Hastings (Metropolis et al., 1953) e o Amostrador de Gibbs (Geman e Geman, 1984) os métodos mais utilizados para tal finalidade. O Amostrador de Gibbs, em vez de gerar amostras diretamente da distribuição conjunta *a posteriori*, transforma o problema multivariado em uma sequência de problemas univariados ao amostrar iterativamente das distribuições condicionais completas *a posteriori*. Essa abordagem simplifica a simulação, especialmente em casos onde a distribuição conjunta é difícil de amostrar diretamente, mas as distribuições condicionais são conhecidas e mais fáceis de manipular. Assim, o método possibilita obter amostras da distribuição conjunta a partir dessas simulações univariadas. As cadeias geradas para cada parâmetro devem apresentar características como estacionariedade e convergência para que possam representar as amostras de suas distribuições marginais e realizar inferências sobre os parâmetros de interesse.

No modelo de LCA, as distribuições condicionais completas *a posteriori* para os parâmetros $\boldsymbol{\gamma}$ e $\boldsymbol{\rho}_{k,c}$ e para ζ_{ic} são dadas por (Costa, Amorim e Bispo, 2021):

$$\boldsymbol{\gamma}|\mathbf{Y},\boldsymbol{\zeta},\rho \sim Dirichlet\left(\sum_{i=1}^{N}\varsigma_{i1}+\alpha_1,\cdots,\sum_{i=1}^{N}\varsigma_{iC}+\alpha_C\right) ,$$

$$\boldsymbol{\rho}|\mathbf{Y},\boldsymbol{\zeta},\gamma \sim Dirichlet\left(\sum_{i=1}^{N}\varsigma_{ic}I(y_{ik}=1)+u_{k,1|c},\cdots,\sum_{i=1}^{N}\varsigma_{ic}I(y_{ik}=R_k)+u_{k,R_k|c}\right) ,$$

$$\boldsymbol{\zeta}|\mathbf{Y},\boldsymbol{\gamma},\boldsymbol{\rho} \propto Multinomial\left(1, \frac{\gamma_1 \prod_{k=1}^{K}\prod_{r_k=1}^{R_k}\rho_{k,r_k|c}^{I(y_{ik}=r_k)}}{\sum_{c=1}^{C}\gamma_c\prod_{k=1}^{K}\prod_{r_k=1}^{R_k}\rho_{k,r_k|c}^{I(y_{ik}=r_k)}}, \cdots, \frac{\gamma_c \prod_{k=1}^{K}\prod_{r_k=1}^{R_k}\rho_{k,r_k|c}^{I(y_{ik}=r_k)}}{\sum_{c=1}^{C}\gamma_c\prod_{k=1}^{K}\prod_{r_k=1}^{R_k}\rho_{k,r_k|c}^{I(y_{ik}=r_k)}}\right),$$

para $c = 1, \ldots, C$.

Deste modo, o algoritmo do amostrador de Gibbs pode ser descrito de acordo com o Algoritmo (1) (Costa, Amorim e Bispo, 2021).

Algoritmo 1: Amostrador de Gibbs para o submodelo de mensuração

Data: $\{\mathbf{Y}_i(.)\}$, $i = 1, \ldots, N$

Result: $P(\boldsymbol{\gamma}, \boldsymbol{\rho}, \boldsymbol{\zeta}|\mathbf{Y})$

Defina o número máximo de iterações T;

Atribua valores iniciais $(\boldsymbol{\gamma}^{(0)}, \boldsymbol{\rho}^{(0)})$

for $t := 1$ ***até*** T **do**

  Realize $\boldsymbol{\zeta}_i^{(t)} \sim P(\boldsymbol{\zeta}_i|\mathbf{Y}, \boldsymbol{\gamma}^{(t-1)}, \boldsymbol{\rho}^{(t-1)})$, para $i = 1, \ldots, N$;

  Realize $\boldsymbol{\gamma}^{(t)} \sim P(\boldsymbol{\gamma}|\mathbf{Y}, \boldsymbol{\zeta}^{(t)}, \boldsymbol{\rho}^{(t-1)})$;

  Realize $\boldsymbol{\rho}_{k,c}^{(t)} \sim P(\boldsymbol{\rho}_{k,c}|\mathbf{Y}, \boldsymbol{\zeta}^{(t)}, \boldsymbol{\gamma}^{(t)})$, para $k = 1, \ldots, K$ e $c = 1, \ldots, C$;

2.4.2.2 Monte Carlo Hamiltoniano

Além do Amostrador de Gibbs, um outro método computacional que pode ser utilizado para estimação bayesiana neste contexto é o Monte Carlo Hamiltoniano (HMC, *Hamiltonian Monte Carlo*, em inglês) (Brooks et al., 2011; Duane et al., 1987; Neal, 2011). O HMC utiliza as derivadas da função de densidade amostrada para gerar transições eficientes que abrangem a distribuição a *posteriori* (Stan Development Team, 2024b). Uma simulação aproximada de dinâmica Hamiltoniana é realizada baseada em integração numérica, que é então corrigida pela realização de uma etapa de aceitação de *Metropolis*. Isto é, o algoritmo introduz variáveis auxiliares de momento, $\boldsymbol{r}$, aos parâmetros $\boldsymbol{\theta}$ da distribuição alvo, e baseia-se em uma densidade conjunta da forma:

$$P(\boldsymbol{r}, \boldsymbol{\theta}) = P(\boldsymbol{r}|\boldsymbol{\theta})P(\boldsymbol{\theta}).$$

Em muitas aplicações do HMC, incluindo o *software* STAN, a densidade auxiliar é uma normal multivariada que não depende dos parâmetros $\boldsymbol{r}$, tal que $\boldsymbol{r} \sim \mathcal{N}(0, M)$, em que M representa a métrica euclidiana (Stan Development Team, 2024b). Este procedimento pode ser visto como uma transformação do espaço paramétrico que torna a amostragem mais eficiente. Após especificar a densidade condicional dos momentos, a densidade conjunta $P(\boldsymbol{r}, \boldsymbol{\theta})$ define um Hamiltoniano $H(\boldsymbol{r}, \boldsymbol{\theta})$ da seguinte forma:

$$
\begin{aligned}
H(\boldsymbol{r}, \boldsymbol{\theta}) &= -\log P(\boldsymbol{r}, \boldsymbol{\theta}) \\
&= -\log P(\boldsymbol{r}|\boldsymbol{\theta}) - \log P(\boldsymbol{\theta}) \\
&= T(\boldsymbol{r}|\boldsymbol{\theta}) + V(\boldsymbol{\theta}),
\end{aligned}
$$

em que os termos $T(\boldsymbol{r}|\boldsymbol{\theta}) = -\log P(\boldsymbol{r}|\boldsymbol{\theta})$, correspondente à densidade sobre a variável auxiliar, e $V(\boldsymbol{\theta}) = -\log P(\boldsymbol{\theta})$, correspondente à densidade do parâmetro de interesse, são conhecidos, respectivamente, como energia cinética e energia potencial (Betancourt e Girolami, 2015; Stan Development Team, 2024b).

Partindo do valor atual dos parâmetros $\boldsymbol{\theta}$, uma transição para um novo estado é gerada em duas etapas antes de ser submetida a uma etapa de aceitação de *Metropolis*. A função hamiltoniana $H(\boldsymbol{r}, \boldsymbol{\theta})$ gera uma transição para um novo estado, primeiro amostrando o momento auxiliar $\boldsymbol{r} \sim \mathcal{N}(0, M)$. Em seguida, o sistema conjunto composto pelos valores atuais dos parâmetros $\boldsymbol{\theta}$ e o novo momento $\boldsymbol{r}$ é avaliado através das equações de Hamilton. Ver mais detalhes em (Betancourt e Girolami, 2015; Stan Development Team, 2024b).

2.4.3 Estimação bayesiana em análise de classes latentes com desfecho distal contínuo

Em Costa, Amorim e Bispo (2021) são descritos quatro procedimentos para estimação do efeito de uma variável latente ζ em uma resposta distal Z usando o modelo de regressão linear. O primeiro método, chamado de Máxima Probabilidade Bayesiana (BMP), classifica cada indivíduo na classe latente com a maior probabilidade posterior estimada através da Equação (2.4). O segundo método, Pseudo-Classe Bayesiana (BPC), utiliza como regra de atribuição para a variável latente vinte valores gerados da probabilidade posterior de ζ. Já o terceiro procedimento, denominado de Método Simplificado Bayesiano (BSM), atribui um valor a ζ através da moda de sua distribuição marginal a posteriori. Após realizado o procedimento de atribuição, esses três métodos incluem a variável latente como uma covariável observada no submodelo estrutural em que Z é o desfecho. Assim, esses métodos são considerados como uma abordagem em duas etapas.

Já o quarto procedimento, denominado de Método Bayesiano Simultâneo (BS), incorpora o erro de medida no submodelo estrutural (Costa, Amorim e Bispo, 2021). Para isso, todos os parâmetros são estimados simultaneamente via amostrador de Gibbs (Li et al., 2018; A. White e Murphy, 2014), e portanto, é considerado uma abordagem em uma única etapa.

2.4.3.1 Submodelo estrutural

Vários modelos estatísticos podem ser considerados para avaliar a relação funcional entre uma variável dependente Z com a variável latente ζ (independente), permitindo a inclusão de outras covariáveis explicativas $\mathbf{X}$ que não são latentes. Para fins de simplicidade, Costa, Amorim e Bispo (2021) consideraram um modelo de regressão linear definido como:

$$\begin{aligned} Z_i = \tau_0 + \tau_1\zeta_{i1} + \tau_2\zeta_{i2} + \cdots + \tau_{C-1}\zeta_{i,C-1} \\ + \tau_C X_{i1} + \tau_{C+1}X_{i2} + \cdots + \tau_P X_{i,P-C+1} + \epsilon_i, \end{aligned} \tag{2.35}$$

em que $\epsilon_i \sim N(0, \sigma^2)$. Isto implica que $\boldsymbol{Z}|\boldsymbol{\zeta}, \mathbf{X}, \tau, \phi \sim N(\tilde{\mathbf{X}}\tau; \phi^{-1})$, com $\phi = \sigma^{-2}$, $\boldsymbol{\tau} = (\tau_0, \ldots, \tau_P)'$ e $\tilde{\mathbf{X}} = (\mathbf{1}, \boldsymbol{\zeta}, \mathbf{X})'$ denota o vetor de variáveis explicativas. O objetivo neste caso é estimar o vetor de parâmetros do submodelo estrutural (τ, ϕ) para permitir a interpretação típica das estimativas dos modelos lineares generalizados (MLG). Se ζ for considerada uma variável observada, são utilizados métodos de classificação-análise e os procedimentos de inferência para os parâmetros do submodelo estrutural são os mesmos dos MLGs usuais.

Ao considerar, por exemplo, uma priori não informativa típica, $P(\tau, \phi) \propto \phi$, a distribuição *a posteriori* condicional para $\tau|\phi$ é multivariada normal com média $\hat{\tau} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'Z$ e matriz de covariância $\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\phi^{-1}$, e a distribuição *a posteriori* marginal para ϕ é Gamma$\left(\frac{N-P-1}{2}, \frac{(Z-\tilde{\mathbf{X}}\hat{\tau})'(Z-\tilde{\mathbf{X}}\hat{\tau})}{2}\right)$ (Costa, Amorim e Bispo, 2021).

Costa, Amorim e Bispo (2021) definem as etapas necessárias para implementação do amostrador de Gibbs para o submodelo estrutural, incluindo ζ como uma variável observada (Algoritmo 2).

Algoritmo 2: Amostrador de Gibbs para o submodelo estrutural com $\boldsymbol{\zeta}$ como covariável observada

Data: $\{\zeta_i(\cdot), Z_i(\cdot), \mathbf{X}_i(\cdot)\}$, $i = 1, \ldots, N$
Result: $P(\boldsymbol{\tau}, \phi|\mathbf{Z}, \boldsymbol{\zeta}, \mathbf{X})$

```
Defina o número máximo de iterações T;
Atribua o valor inicial de φ^(0)
for t := 1 até T do
    Realize τ^(t) ~ P(τ|Z, ζ, X, φ^(t−1));
    Realize φ^(t) ~ P(φ|Z, ζ, X, τ^(t));
```

2.4.3.2 Métodos de duas etapas com estimação bayesiana

As abordagens do tipo classificação-análise para modelar respostas distais geralmente são realizadas em dois passos (Costa, Amorim e Bispo, 2021):

1. Os parâmetros do modelo de LCA são estimados conforme descrito na Equação (2.32) e, através da distribuição *a posteriori* da variável latente ζ obtida nesse passo, os indivíduos são classificados como pertencentes a uma das classes latentes preditas usando alguma regra de classificação.

2. A variável latente ζ é considerada como covariável observada no modelo de regressão para a resposta distal, e a estimação dos parâmetros no submodelo estrutural é realizada.

Através de estudos de simulação, Costa, Amorim e Bispo (2021) observaram que o método BPC apresenta um alto custo computacional, além de desempenho semelhante ou inferior aos métodos BMP, BSM e BS. Além disto, os métodos BMP e BSM utilizados para estimação em duas etapas (*naive*) apresentaram desempenho semelhante.

2.4.3.3 Método de uma etapa com estimação bayesiana

De acordo com Costa, Amorim e Bispo (2021), no processo de estimação em uma etapa todos os parâmetros são estimados simultaneamente por meio de um amostrador de Gibbs. No método BS a variável latente $\boldsymbol{\zeta}$ é incluída diretamente no modelo de regressão em cada iteração do algoritmo, considerando a distribuição *a posteriori* conjunta dos parâmetros $\boldsymbol{\gamma}, \boldsymbol{\rho}, \tau, \phi$ e a variável latente $\boldsymbol{\zeta}$ condicional aos dados $\mathbf{Z}, \mathbf{Y}$ e $\mathbf{X}$, ou seja:

$$P(\gamma, \rho, \boldsymbol{\zeta}, \tau, \phi | \mathbf{Y}, \mathbf{Z}, \mathbf{X}) \propto P(\gamma, \rho, \boldsymbol{\zeta}, \tau, \phi, \mathbf{Y}, \mathbf{Z}, \mathbf{X}) = \\ P(\mathbf{Y}, \boldsymbol{\zeta} | \gamma, \rho) P(\mathbf{Z} | \mathbf{X}, \boldsymbol{\zeta}, \tau, \phi) P(\gamma) P(\rho) P(\tau) P(\phi).$$

Assim, ao utilizar simultaneamente a distribuição posterior dos parâmetros do submodelo de mensuração e estrutural, os erros de medição associados à LCA são incorporados no procedimento de estimação dos parâmetros para o modelo de resposta distal, minimizando as incertezas e atenuações observadas em abordagens tradicionais em duas etapas.

Capítulo 3

Análise de Sobrevivência e Respostas Distais

A análise de sobrevivência é uma área extremamente rica da estatística, com uma ampla variedade de metodologias e aplicações em áreas como engenharia, economia, ciências sociais e medicina. O objetivo central na área de análise de sobrevivência é descrever o tempo de duração dos eventos ou mais comumente, na área médica, descrever o tempo de sobrevivência ou o tempo até a ocorrência de um evento de interesse (também denominado *tempo de falha*) a partir de um marco pré-estabelecido, como, por exemplo, o diagnóstico de uma doença ou a realização de uma cirurgia ou tratamento, em função de fatores de risco (ou variáveis independentes) (Colosimo e Giolo, 2006; Therneau e Grambsch, 2000).

Neste capítulo são introduzidos conceitos básicos da área de análise de sobrevivência (Seção 3.1). Na Seção 3.2, por sua vez, é apresentado um resumo dos aspectos mais relevantes da estimação bayesiana em análise de sobrevivência. Os procedimentos em duas ou três etapas com LCA e o modelo distal com desfecho definido pelo tempo até o evento em dados censurados na abordagem frequentista são descritos na Seção 3.3.

3.1 Análise de sobrevivência

O termo análise de sobrevivência refere-se classicamente a metodologias usualmente consideradas na análise de dados censurados na área da saúde. Entretanto, condições similares ocorrem em outras áreas. Em engenharia, são comuns os estudos em que produtos ou componentes são colocados sob teste para se estimar características relacionadas aos seus tempos de vida. O mesmo ocorre em ciências sociais, em que várias situações de interesse têm como resposta o tempo entre eventos (Colosimo e Giolo, 2006). Nestes contextos, o tempo de início do estudo deve ser precisamente definido, pois os indivíduos

devem ser comparáveis na origem do estudo. Em um estudo clínico aleatorizado, a data da aleatorização ao tratamento é a escolha natural para a origem do estudo. Em estudos observacionais, a data de um diagnóstico ou do início do tratamento de doenças também são outras escolhas possíveis. A escala de medida é quase sempre o tempo real ou "tempo-calendário", apesar de existirem outras alternativas. Em testes de engenharia podem surgir outras escalas de medida, como o número de ciclos, a quilometragem de um carro ou qualquer outra medida de carga.

Vários autores destacam que a principal característica de dados de sobrevivência é a presença de censura, que é a observação parcial da resposta, que ocorre por interrupção no acompanhamento do indivíduo, ou truncamento. A interrupção pode ocorrer porque o indivíduo mudou-se, por término do tempo do estudo ou por morte do indivíduo por causa diferente da estudada (Collett, 1994; Colosimo e Giolo, 2006; Klein e Moeschberger, 2003). Sem a presença de censura, as técnicas estatísticas clássicas, como análise de regressão e planejamento de experimentos, poderiam ser utilizadas na análise deste tipo de dados, provavelmente usando uma transformação para a resposta, que geralmente apresenta assimetria positiva. Já o truncamento ocorre quando há exclusão de certos indivíduos do estudo que não experimentaram algum evento específico (Colosimo e Giolo, 2006).

A censura pode ser classificada em tipos, sendo os mais comuns a censura à esquerda, censura intervalar e censura à direita. A censura à esquerda ocorre quando não se sabe o tempo de ocorrência do evento, mas há o conhecimento de que o mesmo ocorreu antes do tempo registrado. Na censura intervalar sabe-se apenas que o evento de interesse ocorreu em um certo intervalo de tempo, ocorrendo, por exemplo, em estudos em que os pacientes são acompanhados em visitas periódicas e é conhecido somente que o evento de interesse ocorreu em um certo intervalo de tempo. A censura à direita, por sua vez, ocorre quando o tempo de falha não é observável para todos os indivíduos até o final do período de estudo. Assim, esses tempos são denominados censurados à direita, pois o tempo de falha do indivíduo está além, ou seja, é posterior ao seu tempo de censura registrado (Klein et al., 2013). O desprezo dessa informação enviesa o risco estimado de ocorrência do evento de interesse uma vez que o tempo até a falha é desconhecido pois o evento de interesse não ocorreu até o último momento observado (Klein et al., 2013).

De acordo com Klein e Moeschberger (2003), Colosimo e Giolo (2006) e Klein et al. (2013), existem muitos tipos de mecanismos de censura à direita, sendo os mais comuns os tipos I, II ou aleatória. A censura à direita com mecanismo do tipo I ocorre quando alguns pacientes não experimentaram o evento de interesse até o final do estudo. A censura à direita com mecanismo do tipo II ocorre quando o estudo é finalizado após a ocorrência de um número pré-estabelecido de falhas. Já à censura a direita com mecanismo aleatório

ocorre quando o acompanhamento de alguns pacientes foi interrompido por alguma razão não relacionada ao processo em avaliação e outros pacientes não experimentaram o evento até o final do estudo (Colosimo e Giolo, 2006). Nesta dissertação, são apresentadas as definições metodológicas para o contexto de dados censurados à direita, quaisquer que sejam os mecanismos de censura.

3.1.1 Funções de Sobrevivência e Taxa de Falha

Segundo Collett (1994), ao resumir os dados de sobrevivência, existem duas funções de interesse central, que são a função de sobrevivência e a função taxa de falha (ou de risco). Para T uma variável aleatória não-negativa, que representa o tempo de falha, a função de sobrevivência $S(t)$ é definida como a probabilidade de uma observação não falhar até um certo tempo t, ou seja, a probabilidade de que um indivíduo sobreviva ou que o evento ocorra após o tempo t. Esta função é expressa por:

$$S(t) = P(T \geq t). \tag{3.1}$$

A função de sobrevivência $S(t)$ também pode ser definida em termos da distribuição acumulada, uma vez que $F(t) = 1 - S(t)$, que representa a probabilidade de que o tempo de sobrevivência seja menor que algum valor t. A função taxa de falha, por sua vez, é definida por:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \left\{ \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \right\} \\ &= \lim_{\Delta t \to 0} \left\{ \frac{F(t + \Delta t) - F(t)}{\Delta t} \right\} \frac{1}{S(t)} \ . \end{aligned} \tag{3.2}$$

Assumindo Δt muito pequeno, $h(t)$ representa a taxa de falha instantânea no tempo t, condicional à sobrevivência até o tempo t. É importante observar que as taxas de falha são números positivos, mas não possuem um limite superior. Isso implica que a taxa de falha pode variar amplamente e não há um valor máximo para a taxa de falha em um determinado momento no tempo.

Da Equação (3.2) seguem as relações:

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} = -\frac{\partial}{\partial t}(\log(S(t)) \ , \\ S(t) &= \exp\{-H(t)\} \ , \\ H(t) &= -\log(S(t)) = \int_0^t h(q)dq \ , \end{aligned} \tag{3.3}$$

em que $f(t)$ representa a função de densidade de probabilidade da variável aleatória associada ao tempo de sobrevivência T, enquanto a função $H(t)$ é a função taxa de falha acumulada e é mais amplamente discutida em métodos de estimação não paramétricos.

3.1.2 Modelo semiparamétrico de Cox

Em análise de sobrevivência, o modelo de riscos proporcionais de Cox (Cox, 1972) tem sido largamente utilizado em várias áreas do conhecimento para avaliar o efeito de diversas variáveis no risco de desenvolvimento de eventos de interesse, usando-se a informação do tempo em que os eventos ocorreram. O modelo de Cox modela a função taxa de falha em função de uma taxa de falha basal $h_0(t)$ e das covariáveis $\{x_1, \cdots, x_P\}$, de modo que um coeficiente positivo associado aos parâmetros do modelo indica um maior risco de ocorrência do evento. Este modelo é definido através da Equação (3.4):

$$\begin{aligned} h(t|\mathbf{x}) &= h_0(t) \exp\left(\beta_1 x_1 + \cdots + \beta_p x_P\right) \\ &= h_0(t) \exp\left(\mathbf{x}'\boldsymbol{\beta}\right) , \end{aligned} \tag{3.4}$$

em que $\eta = \mathbf{x}'\boldsymbol{\beta}$ é o chamado preditor linear do modelo. A razão entre taxas de falha associadas com um preditor específico pode ser obtida pelo exponencial do correspondente parâmetro. A estimação dos parâmetros do modelo de Cox pode ser feita via método de máxima verossimilhança parcial, maximizando a expressão definida na Equação (3.5):

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{\exp\left(\mathbf{s}'_i\boldsymbol{\beta}\right)}{\left[\sum_{j \in R(t_i)} \exp\left(\mathbf{x}'_j\boldsymbol{\beta}\right)\right]^{d_i}} , \tag{3.5}$$

em que $\mathbf{s}_i$ é o vetor formado pela soma das correspondentes p covariáveis para os indivíduos que falharam no mesmo tempo t_i; $i = 1, \ldots, m$, representa o índice relacionado ao número $m \leq N$ de falhas distintas nos tempos $t_1 \leq t_2 \leq \cdots \leq t_m$, $R(t_i)$ o conjunto dos índices das observações sob risco no tempo t_i e d_i o número de falhas neste mesmo tempo.

O modelo de Cox assume que a razão das taxas de falha de dois diferentes indivíduos é constante ao longo do tempo, sendo, portanto, conhecido como modelo de riscos proporcionais. Esta suposição é geralmente avaliada através dos resíduos de Schoenfeld (Colosimo e Giolo, 2006).

3.2 Estimação bayesiana em análise de sobrevivência

Duas maneiras comuns de modelar dados de sobrevivência são: (a) a modelagem da taxa instantânea de ocorrência do evento como uma função do tempo usando os modelos de riscos proporcionais (conforme descrito na subseção 3.1.2), e (b) modelar diretamente o tempo até o evento ocorrer, usando os modelos de tempo de falha acelerado (AFT) (Brilleman et al., 2020). Segundo estes autores, a maioria dos desenvolvimentos metodológicos que estão implementados em *softwares* para modelagem em análise de sobrevivência é baseada em métodos de estimação de máxima verossimilhança

ou verossimilhança parcial, especialmente devido à popularidade do modelo de Cox. As abordagens Bayesianas, apesar de seus benefícios como a capacidade de fazer afirmações probabilísticas sobre parâmetros, têm recebido menos atenção nesse contexto. No entanto, a inferência bayesiana pode ser tão aplicável à análise de sobrevivência quanto em outras áreas da modelagem estatística.

A abordagem Bayesiana apresenta diversas vantagens em comparação com a abordagem frequentista na análise de sobrevivência, uma vez que modelos de sobrevivência geralmente são difíceis de ajustar através da abordagem frequentista, especialmente na presença de esquemas complexos de censura (Ibrahim, Chen e Sinha, 2001). Esta abordagem simplifica esse processo ao empregar amostradores de Gibbs e MCMC, permitindo um ajuste mais direto de modelos complexos e facilitando a implementação por meio de ferramentas como o WinBUGS, JAGS e STAN, utilizadas no ajuste de modelos em análise de sobrevivência (Chen et al., 2014; Işık, Karasoy e Karabey, 2023).

A estimação bayesiana em análise de sobrevivência também oferece uma abordagem mais direta para a inferência de parâmetros de interesse, diferentemente da abordagem frequentista, que muitas vezes depende de resultados assintóticos. A estimação Bayesiana baseada em MCMC possibilita a obtenção de inferências exatas para qualquer tamanho de amostra. Além disso, estimativas de variâncias dos parâmetros do modelo, assim como qualquer outro resumo posterior, são obtidas como um subproduto do amostrador de Gibbs e, portanto, são triviais de se obter uma vez que as amostras da distribuição posterior estão disponíveis (Ibrahim, Chen e Sinha, 2001).

Considerando um contexto em que haja dados de sobrevivência com censura à direita ou com censura intervalar, descritas na seção (3.1), a análise bayesiana semiparamétrica para dados de sobrevivência censurados à direita, baseada em qualquer modelo para a função de sobrevivência $S(t|\ddot{\mathbf{x}};\boldsymbol{\varrho})$, com $\boldsymbol{\varrho}$ representando o conjunto dos parâmetros do modelo e $\ddot{\mathbf{x}}$ um vetor de covariáveis, pode ser conduzida usando a função de verossimilhança para tempos de sobrevivência contínuos ou discretizados (Klein et al., 2013). Nas duas próximas subseções são apresentadas as duas abordagens mais comuns para estimação bayesiana em análise de sobrevivência com dados censurados à direita.

3.2.1 Estimação bayesiana no modelo semiparamétrico de risco constante fragmentado

Conforme Klein et al. (2013), para uma função $f(t|\ddot{\mathbf{x}};\boldsymbol{\varrho})$ representando a densidade sob o modelo usado para descrever a função de sobrevivência $S(t|\ddot{\mathbf{x}};\boldsymbol{\varrho})$, a verossimilhança considerando dados observados contínuos censurados à direita $\boldsymbol{\mathcal{D}} = (N, \mathbf{U}, \boldsymbol{\nu}, \ddot{\mathbf{X}})$ é dada

pela Equação (3.6):

$$L(\boldsymbol{\varrho}|\boldsymbol{\mathcal{D}}) = \prod_{i=1}^{N} f(u_i|\ddot{\mathbf{x}}_i;\boldsymbol{\varrho})^{\nu_i} S(u_i|\ddot{\mathbf{x}}_i;\boldsymbol{\varrho})^{1-\nu_i} \ , \tag{3.6}$$

com $\mathbf{U} = (u_1, u_2, \ldots, u_N)'$, $\boldsymbol{\nu} = (\nu_1, \nu_2, ..., \nu_N)'$, em que ν_i é o indicador de falha, u_i é o tempo de sobrevivência observado sujeito à censura à direita, e $\ddot{\mathbf{X}}$ é a matriz $N \times P$ de covariáveis com a i-ésima linha $\ddot{\mathbf{x}}_i'$.

Segundo Gelfand e Mallick (1995), Ibrahim, Chen e Sinha (2001) e Klein et al. (2013), quando o modelo de sobrevivência é descrito pelo modelo de Cox, apresentado na Equação (3.4), com covariáveis fixas $\ddot{\mathbf{x}}$, a verossimilhança definida em (3.6) pode ser expressa como:

$$L(\boldsymbol{\varrho}|\boldsymbol{\mathcal{D}}) = \prod_{i=1}^{N} \exp\left\{ -H_{0(i)} \sum_{j \in R(i)} \exp\left(\ddot{\mathbf{x}}_j'\boldsymbol{\varphi}\right) \right\} \left\{ h_0(u_{(i)}) \exp\left(\ddot{\mathbf{x}}_{(i)}'\boldsymbol{\varphi}\right) \right\}^{\nu_{(i)}} \ , \tag{3.7}$$

com $\boldsymbol{\varrho} = (\boldsymbol{\varphi}, H_0(\cdot))$, em que $H_0(t)$ denota a taxa de risco acumulada basal, $\boldsymbol{\varphi}$ é o vetor de parâmetros de interesse, $0 = u_{(0)} < u_{(1)} < \cdots < u_{(N)}$ são os tempos de sobrevivência observados ordenados, e $\ddot{x}_{(i)}$ as covariáveis associadas a $u_{(i)}$. O termo $H_{0(i)} = H_0(u_{(i)}) - H_0(u_{(i-1)})$ representa o incremento na taxa de risco acumulada entre os tempos de sobrevivência observados consecutivos $u_{(i-1)}$ e $u_{(i)}$ e $R_{(i)} = \{j : u_{(j)} \geq u_{(i)}\}$ é o conjunto de todos os indivíduos sob risco no tempo j.

Embora os métodos bayesianos semiparamétricos possam lidar com a verossimilhança adaptada para tempos discretizados, descrita com mais detalhes em Klein et al. (2013), é comum utilizar a verossimilhança contínua da Equação (3.6) como uma aproximação para sua versão considerando tempos discretizados, desde que a grade de monitoramento não seja muito destoante em relação à escala do tempo de sobrevivência. A verossimilhança de tempo contínuo pode ser vista como um caso limite da verossimilhança de tempo discreto, e existe uma percepção geral de que não há muitas vantagens em utilizar $h_0(t)$ estritamente não paramétrico juntamente com a verossimilhança de tempo discreto da Equação (3.6).

Nesta abordagem, um dos modelos mais convenientes e populares para análise de sobrevivência semi-paramétrica é o modelo de risco constante por partes, também conhecido como fragmentado (ou *piecewise*) (Ibrahim, Chen e Sinha, 2001; Klein et al., 2013). Para construção deste modelo, inicialmente é criada uma partição finita do eixo do tempo, $0 < s_1 < s_2 < \cdots < s_J$, com $s_J > u_i$ para todo $i = 1, 2, \ldots, N$. Assim, tem-se os J intervalos $(0, s_1], (s_1, s_2], \ldots, (s_{J-1}, s_J]$. No j-ésimo intervalo, é assumido um risco de base constante $h_0(y) = \lambda_j$ para $y \in I_j = (s_{j-1}, s_j]$. Definindo $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_J)$,

pode-se escrever a função de verossimilhança de $(\boldsymbol{\varphi}, \boldsymbol{\lambda})$ para os N indivíduos como:

$$L(\boldsymbol{\varphi}, \boldsymbol{\lambda}|\boldsymbol{\mathcal{D}}) = \prod_{i=1}^{N}\prod_{j=1}^{J}(\lambda_j \exp(\ddot{\mathbf{x}}_i'\boldsymbol{\varphi}))^{\delta_{ij}\nu_i} \times \exp\left\{-\delta_{ij}\left[\lambda_j(u_i - s_{j-1}) + \sum_{g=1}^{j-1}\lambda_g(s_g - s_{g-1})\right]\exp(\ddot{\mathbf{x}}_i'\boldsymbol{\varphi})\right\}, \tag{3.8}$$

com $\delta_{ij} = 1$ se o i-ésimo indivíduo falhou ou foi censurado no j-ésimo intervalo, e 0 caso contrário. O indicador δ_{ij} é necessário para definir corretamente a verossimilhança sobre os J intervalos para o modelo semiparamétrico apresentado na Equação (3.8), que também pode ser denominado "modelo exponencial por partes", permitindo acomodar várias formas do risco basal nos intervalos (Klein et al., 2013).

De acordo com Ibrahim, Chen e Sinha (2001), o modelo se reduz a um modelo exponencial paramétrico com parâmetro de taxa de falha $\lambda \equiv \lambda_1$ para o caso em que $J = 1$. Uma priori comum para o risco basal λ é gama independente, tal que $\lambda_j \sim \mathrm{G}(\alpha_{0j}, \lambda_{0j})$ para $j = 1, 2, \ldots, J$, com hiperparâmetros *a priori* α_{0j} e λ_{0j}, que podem ser obtidos através da média e da variância *a priori* de λ_j (Ibrahim, Chen e Sinha, 2001; Klein et al., 2013). Outra metodologia mencionada na literatura é o estabelecimento de uma correlação *a priori* entre os $\lambda_j's$ (Leonard, 1978; Sinha, 1993), de modo que seja usada uma distribuição *a priori* $\boldsymbol{\psi} \sim \mathcal{N}(\psi_0, \Sigma_\psi)$, onde $\psi_j = \log(\lambda_j)$ para $j = 1, 2, \ldots, J$.

3.3 Respostas distais em análise de sobrevivência

A análise de classes latentes com desfecho distal do tipo tempo até o evento busca compreender a relação entre variáveis latentes e a ocorrência de um evento ao longo do tempo. Dois artigos relevantes nesse contexto são os artigos de Larsen (2004) e Lythgoe, Garcia-Fiñana e Cox (2019). O primeiro artigo apresenta a extensão do modelo de classes latentes, com indicadores binários e preditores das classes latentes, a fim de incorporar a variável latente categórica como preditora de uma resposta distal de tempo até o evento, utilizando o procedimento de estimação em uma etapa (Larsen, 2004). Já o segundo artigo compara diferentes abordagens de modelagem de classes latentes, considerando indicadores politômicos, com um desfecho distal de tempo até o evento, analisando os benefícios e limitações das abordagens de uma, duas e três etapas, sendo consideradas também as abordagens de três etapas denominadas inclusivas, que tratam o desfecho distal inicialmente como variável preditora para a variável latente (Lythgoe, Garcia-Fiñana e Cox, 2019).

A seguir são apresentados os métodos de estimação em 1, 2 e 3 etapas já desenvolvidos, sob a ótica frequentista, para estimação em modelos com desfecho distal de tempo

até o evento.

3.3.1 Estimação frequentista em uma etapa em LCA com desfecho distal de tempo até o evento

Para estimação em uma etapa existem duas abordagens disponíveis na literatura. A primeira abordagem foi desenvolvida em Larsen (2004) definindo um modelo que comporta indicadores binários e covariáveis preditoras da variável latente categórica, e um desfecho distal de tempo até o evento, que pode ter outras covariáveis preditoras. O modelo apresentado por Lythgoe, Garcia-Fiñana e Cox (2019), por outro lado, comporta indicadores politômicos para a análise de classes latentes.

Para este tipo de abordagem, considere os vetores $\mathbf{Y}_i = (Y_{i1}, \cdots, Y_{iK})'$ e $\mathbf{X}_i = (X_{i1}, \cdots, X_{iQ})'$ definidos, respectivamente, como um vetor aleatório de K indicadores e um vetor linha $1 \times Q$ de preditores da variável latente ζ. E seja $\ddot{\mathbf{x}}_i = (\ddot{x}_{i1}, \cdots, \ddot{x}_{iP})'$ um vetor $1 \times P$ de covariáveis observadas discretas ou contínuas do i-ésimo indivíduo. O modelo de riscos proporcionais assume que a função taxa de falha para o tempo de evento do i-ésimo indivíduo é da forma:

$$\begin{aligned} h(t|\ddot{\mathbf{x}}, \zeta) &= \lim_{h \to 0^+} \left(\frac{1}{h}\right) P(T_i < t + h | T_i \geq t, \ddot{\mathbf{x}}_i, \zeta = c) \\ &= h_i(t|\ddot{\mathbf{x}}_i, \zeta = c) \\ &= h_0(t) \exp(\ddot{\mathbf{x}}_i \boldsymbol{\varphi} + \beta_c) \ , \end{aligned} \tag{3.9}$$

em que $\boldsymbol{\varphi}$ é o vetor $P \times 1$ de parâmetros associados a $\ddot{\mathbf{x}}_i$, e $\boldsymbol{\beta}$ é o vetor $C \times 1$ de parâmetros contendo os efeitos da variável de classe latente ζ sobre o risco, com $\beta_1 = 0$ para fins de identificação. A forma da taxa de falha basal $h_0(t)$ não é especificada, sendo estimada não parametricamente. Daí, usando as relações definidas em (3.3), a densidade da distribuição do tempo do evento T_i pode ser expressa por:

$$P(t|\ddot{\mathbf{x}}_i, \zeta = c) = h_0(t) \exp(\ddot{\mathbf{x}}_i \boldsymbol{\varphi} + \beta_c) \exp\{-H_0(t) \exp(\ddot{\mathbf{x}}_i \boldsymbol{\varphi} + \beta_c)\} \ ,$$

em que $H_0(t) = \int_0^t h_0(s)ds$ é a função taxa de falha acumulada basal, sendo T_i o tempo de falha, que geralmente é censurado à direita. Seja T_i^* o tempo de censura, de modo que as variáveis observáveis são $U_i = \min(T_i, T_i^*)$ e $\nu_i = I(T_i \leq T_i^*)$. No caso de censura não informativa (Andersen et al., 1993), a distribuição de probabilidade de (U_i, ν_i) se torna

$$P(u_i, \upsilon_i|\ddot{\mathbf{x}}_i, \zeta = c) \propto \{h_0(u_i) \exp(\ddot{\mathbf{x}}_i \boldsymbol{\varphi} + \beta_c)\}^{\upsilon_i} \times \exp\{-H_0(u_i) \exp(\ddot{\mathbf{x}}_i \boldsymbol{\varphi} + \beta_c)\} \ , \tag{3.10}$$

em que ν denota o indicador de falha, de modo que υ assume o valor 1 se o evento for observado e 0 caso contrário. Assume-se aqui que a censura é não informativa.

Considerando-se um modelo de LCA com covariáveis para regressão de $(\mathbf{Y}_i, \zeta)$ em $\mathbf{x}_i$ e um modelo de riscos proporcionais para a regressão de (U_i, ν_i) em $(\ddot{\mathbf{x}}_i, \zeta)$, a distribuição conjunta de $(U_i, \nu_i, \mathbf{Y}_i, \zeta)$ será dada por:

$$P(u_i, \upsilon_i, \mathbf{y}_i, \zeta = c|\mathbf{x}_i, \ddot{\mathbf{x}}_i) = P(\zeta = c|\mathbf{x}_i)P(\mathbf{y}_i|\zeta = c)P(u_i, \upsilon_i|\ddot{\mathbf{x}}_i, \zeta = c). \tag{3.11}$$

Ao integrar a Equação (3.11) sobre a variável de classe latente ζ, a distribuição marginal para as variáveis observadas $(U_i, \nu_i, \mathbf{Y}_i)$ pode ser representada como:

$$P(u_i, \upsilon_i, \mathbf{y}_i|\mathbf{x}_i, \ddot{\mathbf{x}}_i) = \sum_{c=1}^{C} P(\zeta = c|\mathbf{x}_i)P(\mathbf{y}_i|\zeta = c)P(u_i, \upsilon_i|\ddot{\mathbf{x}}_i, \zeta = c). \tag{3.12}$$

De acordo com Larsen (2004), o algoritmo EM é utilizado para maximizar a probabilidade dos dados observados $(\mathbf{U}, \boldsymbol{\nu}, \mathbf{Y})$. Esse processo ocorre por meio de iterações entre duas etapas: na etapa E, em que é calculada a log-verossimilhança esperada dos dados completos $(\mathbf{U}, \boldsymbol{\nu}, \mathbf{Y}, \boldsymbol{\zeta})$, com base nos dados observados e na estimativa atual dos parâmetros, e na etapa M, em que novas estimativas dos parâmetros são obtidas ao maximizar a log-verossimilhança esperada.

A log-verossimilhança para estimação dos parâmetros $\dot{\boldsymbol{\varrho}} = (\boldsymbol{\rho}, \boldsymbol{\vartheta}, \mathbf{h}, \boldsymbol{\varphi}, \boldsymbol{\beta})'$ pelo modelo de respostas distais em análise de sobrevivência pode ser expressa por (Larsen, 2004):

$$l(\dot{\boldsymbol{\varrho}}) = \sum_{i=1}^{N} \log \left[\sum_{c=1}^{C} P(\zeta = c|\mathbf{x}_i)P(\mathbf{y}_i|\zeta = c)P(u_i, \upsilon_i|\ddot{\mathbf{x}}_i, \zeta = c) \right], \tag{3.13}$$

em que o vetor de parâmetros $\boldsymbol{\vartheta}$ contido em $\dot{\boldsymbol{\varrho}}$ é relacionado ao termo $P(\zeta = c|\mathbf{x}_i)$ definido no submodelo estrutural (no contexto de LCA com covariáveis), definido na Equação (2.12).

A partir dos desenvolvimentos de Lythgoe, Garcia-Fiñana e Cox (2019), a log-verossimilhança para estimação do vetor de parâmetros $\ddot{\boldsymbol{\varrho}} = (\boldsymbol{\rho}, \boldsymbol{\gamma}, \mathbf{h}, \boldsymbol{\varphi}, \boldsymbol{\beta})'$ pelo modelo de respostas distais em análise de sobrevivência pode ser expressa por:

$$l(\ddot{\boldsymbol{\varrho}}) = \sum_{i=1}^{N} \log \left[\sum_{c=1}^{C} P(\zeta = c)P(\mathbf{y}_i|\zeta = c)f_{T,\nu|\ddot{\mathbf{x}},\zeta}(t, \upsilon|\ddot{\mathbf{x}}, c) \right]. \tag{3.14}$$

Os vetores de parâmetros $\dot{\boldsymbol{\varrho}}$ e $\ddot{\boldsymbol{\varrho}}$ diferem entre as abordagens apresentadas em Larsen (2004) e Lythgoe, Garcia-Fiñana e Cox (2019), respectivamente, pois a abordagem desenvolvida por Larsen (2004) considera a situação em que a variável latente ζ possui preditores observados. Os vetores de parâmetros $\boldsymbol{\gamma}$ e $\boldsymbol{\rho}$ são relacionados, respectivamente, aos termos $P(\zeta = c)$ e $P(\mathbf{y}_i|\zeta = c)$ do submodelo de mensuração no contexto de LCA,

definidos na Equação (2.3).

A abordagem desenvolvida por Lythgoe, Garcia-Fiñana e Cox (2019) modela a função de densidade do tempo até o evento utilizando-se do modelo exponencial segmentado (Friedman, 1982), assumindo que a função de risco basal é segmentada constante. Desta forma o tempo t é particionado em $s = 1, \ldots, S$ intervalos, com $\mathbf{h}_0 = (h_{01}, \ldots, h_{0S})$ representando o vetor de parâmetros de risco basal. A função $f_{T,\nu|\ddot{\mathbf{x}},\zeta}(t, \upsilon|\ddot{\mathbf{x}}, \zeta)$ é então definida como:

$$f_{T,\nu|\ddot{\mathbf{x}},\zeta}(t, \upsilon|\ddot{\mathbf{x}}, \zeta) = \prod_{s=1}^{S} [h_{0s} \exp(\ddot{\mathbf{x}}\boldsymbol{\varphi} + \zeta\boldsymbol{\beta})]^{\upsilon\psi_s} \times \exp\left\{ -\psi_s \left[h_{0s}(t - h_{s-1}) + \sum_{r=1}^{s-1} h_{0r}(h_r - h_{r-1}) \right] \exp(\ddot{\mathbf{x}}\boldsymbol{\varphi} + \zeta\boldsymbol{\beta}) \right\} , \tag{3.15}$$

em que ψ_s é igual a 1 se o evento ocorrer no s-ésimo intervalo e 0 caso contrário, h_s é o limite superior para o s-ésimo intervalo na grade de tempo, e h_0 é igual a 0 (Lythgoe, Garcia-Fiñana e Cox, 2019).

3.3.2 Estimação frequentista em duas e três etapas em LCA com desfecho distal de tempo até o evento

3.3.2.1 Estimação em duas etapas

Em Lythgoe, Garcia-Fiñana e Cox (2019) são definidos os passos para aplicação da abordagem de estimação em duas etapas, desenvolvida em Bakk e Kuha (2017). Os parâmetros do modelo são divididos em dois grupos, a serem estimados nas etapas 1 e 2, sendo representados por $\ddot{\boldsymbol{\varrho}} = (\ddot{\boldsymbol{\varrho}}_1, \ddot{\boldsymbol{\varrho}}_2)$. No primeiro passo, um modelo de classes latentes é ajustado utilizando a Equação (2.3), estimando os parâmetros $\ddot{\boldsymbol{\varrho}}_1 = (\boldsymbol{\gamma}, \boldsymbol{\rho})$, e portanto, $\ddot{\boldsymbol{\varrho}}_2 = (\mathbf{h}, \boldsymbol{\varphi}, \boldsymbol{\beta})$. As estimativas obtidas nesse passo são denotadas por $\underset{\sim}{\ddot{\boldsymbol{\varrho}}}_1$. No segundo passo, o objetivo é maximizar a log-verossimilhança dos dados observados, condicional às estimativas obtidas na etapa 1, ou seja, $l(\ddot{\boldsymbol{\varrho}}_2|\ddot{\boldsymbol{\varrho}}_1 = \underset{\sim}{\ddot{\boldsymbol{\varrho}}}_1)$. A log-verossimilhança necessária é apresentada na Equação (3.14).

É importante destacar que devido aos parâmetros estimados na etapa 1 serem mantidos fixos durante a estimação na etapa 2, as estimativas resultantes na segunda etapa, $\underset{\sim}{\ddot{\boldsymbol{\varrho}}}_2$, tendem a subestimar a incerteza associada a esses parâmetros (Lythgoe, Garcia-Fiñana e Cox, 2019). Por esta razão, Xue e Bandeen-Roche (2004) e Bakk e Kuha (2017) propuseram métodos para obter erros-padrão corrigidos na abordagem de dois passos.

3.3.2.2 Estimação em três etapas

Para realização da estimação em três etapas, utiliza-se um processo sequencial considerando as abordagens (i) usual/padrão ou (ii) inclusiva (Lythgoe, Garcia-Fiñana e Cox, 2019). No processo padrão de três etapas, a etapa 1 consiste em ajustar um modelo de classes latentes, conforme Equação (2.3). Já no processo inclusivo de três etapas, é utilizado um modelo de regressão de classes latentes, no qual a variável resposta distal é incorporada como um preditor das classes latentes, juntamente com outras covariáveis do modelo para o desfecho. No caso particular em que o desfecho distal é um tempo até o evento sujeito a censura, sugere-se a utilização de metodologia proposta em I. White e Royston (2009), que faz o uso da função taxa de falha acumulada estimada, juntamente com o indicador de evento e outras covariáveis relacionadas ao desfecho tempo até evento.

Para ilustrar esta abordagem inclusiva considerando-se apenas uma covariável $\ddot{x}$, o modelo LCA com covariáveis, equivalente à Equação (2.12), expresso na escala *logit*, é definido por:

$$logitP(\zeta = c|H(t), \nu = \upsilon, \ddot{\mathbf{X}} = \ddot{\mathbf{x}}) = \kappa_{c0} + \kappa_{c1}H(t) + \kappa_{c2}\upsilon + \kappa_{c3}z \ ,$$

para $c = 1, \ldots, C$, com $\ddot{\mathbf{X}}$ representando o vetor de covariáveis preditoras do modelo semiparamétrico de Cox, T representando o tempo até ocorrência do evento, com valor observado t e ν denotando o indicador de falha. Já $H(t)$ expressa a função taxa de falha acumulada, que é estimada separadamente, com abordagem não paramétrica pelo método de Nelson-Aalen (Lythgoe, Garcia-Fiñana e Cox, 2019).

Na etapa 2, os indivíduos são geralmente alocados a uma classe latente de acordo com uma regra de alocação, como, por exemplo, a alocação modal (AM), alocação aleatória (PC), múltiplas imputações de pseudo-classe (mPC), alocação parcial (PA) e alocação proporcional (PrA). A regra de alocação mais simples e comumente utilizada é a alocação modal (AM), na qual cada indivíduo é atribuído à classe latente com a maior probabilidade posterior. Na alocação aleatória, também conhecida como método "pseudo-classe"(PC), a classe é imputada a cada indivíduo somente uma vez, sendo escolhida aleatoriamente a partir de uma distribuição multinomial com probabilidades iguais às probabilidades posteriores estimadas na LCA. Na técnica de múltiplas imputações de pseudo-classe (mPC) (C. Wang, Brown e Bandeen-Roche, 2005), por sua vez, a classe é imputada várias vezes para cada indivíduo de modo similar ao método PC, sendo recomendado realizar pelo menos 20 imputações aleatórias. Na alocação parcial (PA), nenhuma atribuição é feita e as probabilidades posteriores são utilizadas em análises adicionais. E, por fim, na alocação proporcional (PrA), cada sujeito é atribuído a todas as classes simultaneamente, com pesos iguais às suas probabilidades específicas das classes correspondentes (Lythgoe, Garcia-Fiñana e Cox, 2019).

Na etapa 3, a variável resposta distal é regredida em função das classes latentes atribuídas da etapa 2 e de outras covariáveis relevantes. No caso em que a resposta distal é o tempo até o evento, geralmente tem-se considerado o ajuste do modelo de riscos proporcionais de Cox (Cox, 1972) nesta etapa. Para os métodos AM e PC, $C-1$ variáveis *dummy* são usadas para representar a classe atribuída/imputada no modelo de regressão. Para o método mPC, esse processo é repetido para cada imputação de classe, e as estimativas dos parâmetros são combinadas entre os modelos de regressão usando as regras de imputação de Rubin (Rubin, 2004). No método PA, $C-1$ probabilidades posteriores são incluídas como covariáveis no modelo de regressão. No método PrA, cada sujeito é incluído no modelo de regressão C vezes, com pesos iguais às probabilidades posteriores do modelo de classe latente. Uma consequência do método PrA em um cenário de tempo até o evento é a introdução de tempos de eventos empatados.

Não foi encontrada na literatura discussão a respeito da estimação dos parâmetros no modelo de resposta distal em análise de sobrevivência usando a abordagem bayesiana. Sendo assim, no Capítulo 4 são apresentadas as metodologias propostas para o ajuste deste modelo, considerando-se as definições apresentadas na revisão de literatura dos Capítulos 2 e 3.

Capítulo 4

Métodos Bayesianos para Respostas Distais em Análise de Sobrevivência

Como apresentado no Capítulo 2, sobretudo nas Seções 2.3 e 2.4.3, existem diversas abordagens, seja frequentista ou bayesiana, que podem ser utilizadas para estimação do efeito distal em um modelo de classes latentes, tanto para desfechos categóricos quanto contínuos. Todavia, quando a variável resposta distal representa um tempo censurado até o evento encontram-se apenas propostas de métodos na abordagem frequentista (Larsen, 2004; Lythgoe, Garcia-Fiñana e Cox, 2019), conforme descritos no Capítulo 3, Seção 3.3. No presente trabalho são propostas duas metodologias bayesianas para estimação do efeito de uma variável latente categórica, proveniente de LCA, no tempo até o evento na análise de dados censurados.

Neste capítulo apresentamos a especificação dos modelos de interesse (Subseção 4.1), seguida das metodologias propostas, que são denominadas, respectivamente, de método Bayesiano Modal Simplificado (BSM) e método Bayesiano Simultâneo (BS) (Subseção 4.2) além de detalhes acerca da implementação das metodologias através de *software* estatístico (Subseção 4.3).

4.1 Especificação dos Modelos

O modelo de interesse consiste em duas partes: um submodelo de mensuração, definido via análise de classes latentes, considerando-se indicadores categóricos, e um submodelo estrutural, baseado no modelo de riscos proporcionais de Cox. Ambos os submodelos são definidos a seguir.

4.1.1 Submodelo de mensuração

Seja C o número de classes da variável latente ζ e K o número de variáveis indicadoras politômicas observadas, em que uma variável observada k possui R_k categorias de resposta. O vetor do padrão de respostas para os N indivíduos do estudo é representado por $\mathbf{Y} = (\mathbf{Y_1}, \mathbf{Y_2}, ..., \mathbf{Y_n})'$ em que $\mathbf{Y_i} = (Y_{i1}, Y_{i2}, ..., Y_{iK})'$ corresponde ao padrão individual de resposta para o i-ésimo indivíduo. Para a definição do submodelo de mensuração, considere que a função de verossimilhança dos dados completos para o i-ésimo indivíduo é definida por:

$$P(\mathbf{Y}_i = \mathbf{y}_i, \boldsymbol{\xi}_i = \boldsymbol{\varsigma}_i | \boldsymbol{\gamma}, \boldsymbol{\rho}) = \prod_{c=1}^{C} \left[\gamma_c \prod_{k=1}^{K} \prod_{r_k=1}^{R_k} \rho_{k,r_k|c}^{I(y_{ik}=r_k)} \right]^{\varsigma_{ic}} , \tag{4.1}$$

em que $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iC})'$ é o vetor de respostas multinomiais relacionado à variável latente para o i-ésimo indivíduo, com valores observados $\boldsymbol{\varsigma}_i = (\varsigma_{i1}, \ldots, \varsigma_{iC})'$, de forma que $\varsigma_{ic} = 1$ se o indivíduo i pertence à classe c e igual a 0 caso contrário. Adicionalmente $P(\xi_{ic} = 1) = \gamma_c$ para $c = 1, \cdots, C$, com $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_n)'$.

No submodelo de mensuração o vetor de parâmetros é definido por $\boldsymbol{\varrho_1} = (\boldsymbol{\gamma}, \boldsymbol{\rho})$, tal que as prevalências das classes latentes são denotadas por $\gamma_c = P(\zeta = c)$, e as probabilidades condicionais de resposta ao item são denotadas por $\rho_{k,r_k|c} = P(Y_{ik} = r_k | \zeta = c)$. Os vetores de parâmetros $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_C)'$ e $\boldsymbol{\rho}_{k,c} = (\rho_{k,1|c}, \cdots, \rho_{k,R_k|c})'$ seguem ambos distribuições a priori de Dirichlet, com vetores de hiperparâmetros $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_C)'$ e $\boldsymbol{\varepsilon}_{k,c} = (\varepsilon_{k,1|c}, \cdots, \varepsilon_{k,R_k|c})$, respectivamente (Costa, Amorim e Bispo, 2021). A função de densidade de probabilidade conjunta a priori para $\boldsymbol{\rho} = (\boldsymbol{\rho}_1, \cdots, \boldsymbol{\rho}_c)'$, em que $\boldsymbol{\rho}_c = (\boldsymbol{\rho}_{1,1}, \cdots, \boldsymbol{\rho}_{K,C})'$ é definida como

$$P(\boldsymbol{\rho}) \propto \prod_{c=1}^{C} \prod_{k=1}^{K} \prod_{r_k=1}^{R_k} \rho_{k,r_k|c}^{\varepsilon_{k,r_k|c}-1} ,$$

enquanto que a função de densidade de probabilidade conjunta a priori para $\boldsymbol{\gamma}$ é dada por:

$$P(\boldsymbol{\gamma}) \propto \prod_{c=1}^{C} \gamma_c^{\alpha_c - 1} .$$

4.1.2 Submodelo estrutural

Seja T_i o tempo de falha, potencialmente censurado à direita, $i = 1, 2, \ldots, N$, e T_i^* o tempo de censura, tal que $U_i = \min(T_i, T_i^*)$ denota o tempo observado e $\nu_i = I(T_i \leq T_i^*)$ é o indicador de falha. O modelo de riscos proporcionais assume que a função taxa de

falha para o tempo de evento do i-ésimo indivíduo é da forma:

$$h(t|\ddot{\mathbf{x}}_i, \zeta_\mathbf{i}) = h_0(t) \exp(\ddot{\mathbf{x}}_i \boldsymbol{\varphi} + \zeta_\mathbf{i} \boldsymbol{\beta}) \ , \tag{4.2}$$

em que $\boldsymbol{\varrho_2} = (\boldsymbol{\varphi}, \boldsymbol{\beta})$ denota o vetor de parâmetros do submodelo estrutural. O vetor $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_c)$ contém os efeitos das classes latentes no risco de ocorrência de um evento na análise de dados censurados, sendo $\beta_1 = 0$ para identificação, enquanto o vetor $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_P)$ contém os efeitos associados às P variáveis preditoras observadas do tempo até ocorrência do desfecho, contidas no vetor de covariáveis $\ddot{\mathbf{X}}$.

De acordo com a proposta inicial de Larsen (2004) para estimação dos efeitos de uma variável latente no tempo até a ocorrência de um evento, o logaritmo da função de verossimilhança para estimação destes parâmetros pode ser expresso segundo Equação (3.13). Na abordagem clássica de estimação, envolvendo o modelo semiparamétrico de riscos proporcionais de Cox, a taxa de falha basal não é considerada diretamente durante a estimação dos parâmetros, que se dá através do método de máxima verossimilhança parcial (Equação (3.5)). Contudo, em sua proposta com abordagem frequentista, Larsen (2004) considera uma estimação não paramétrica para o termo $h_0(t)$ (taxa de falha basal), em que é assumida que a taxa de falha basal é constante ao longo do tempo.

Ao realizar estimação bayesiana em análise de sobrevivência semiparamétrica é comum considerar a taxa de falha basal durante o processo de estimação dos parâmetros, assumindo inclusive prioris para a taxa de falha basal considerando um modelo de risco constante por partes (*piecewise*) (Ibrahim, Chen e Sinha, 2001; Klein et al., 2013). Entretanto, a abordagem de estimação por meio do modelo *piecewise* assume que existem $s = 1, \ldots, S$ intervalos em que o tempo t seja particionado, e em cada um destes intervalos a taxa de falha basal é constante, o que pode ser um pressuposto irreal de acordo com a natureza de muitos fenômenos estudados, como em doenças progressivas (câncer ou infecção por HIV), em que o risco de mortalidade ou agravamento pode aumentar continuamente com o tempo à medida que a condição piora, ou diminuir conforme o tratamento faz efeito.

Desta forma, a abordagem de *M-splines* para o modelo semiparamétrico de Cox é útil, uma vez que, segundo W. Wang e Yan (2021), os *splines* (Curry e Schoenberg, 1966; Ramsay, 1988) são escolhas naturais para modelar a função de risco basal $h_0(t)$ e a função de risco basal cumulativo $H_0(t)$, respectivamente, ou seus logaritmos, no contexto de dados de tempo até o evento. Sleeper e Harrington (1990) mencionam que a resposta no modelo semiparamétrico de Cox é a função taxa de falha, descrita a partir de uma taxa de risco logarítmica que é linear nas covariáveis. Todavia, havendo indícios de violação dessa suposição, os efeitos das covariáveis são melhor representados por funções suaviza-

das (*smooth*) e não lineares. Desta forma, a abordagem com modelos semiparamétricos de sobrevivência utilizando *splines* é indicada como sendo mais flexível, não requerendo linearidade, além de ser útil em conjuntos de dados complexos incluindo a variável tempo de falha censurada. Aplicações deste tipo de metodologia no contexto de modelos semiparamétricos para sobrevivência são encontradas em artigos como o de Rosenberg (1995), que utilizou *B-splines* para modelagem da função de risco basal, o de Brilleman et al. (2020), com seu uso para modelos semiparamétricos de sobrevivência com covariáveis tempo-dependentes, e o de Heinzl e Kaider (1997) com o uso de *splines* cúbicas para obter mais flexibilidade no modelo de riscos proporcionais de Cox (Equação (3.4)).

Considerando o uso de *M-Splines*, também conhecidos como *B-splines* Curry-Schoenberg, definidos por Curry e Schoenberg (1966), os autores Brilleman et al. (2020) definiram uma nova parametrização flexibilizando a classe de modelos de sobrevivência na escala de risco, o que inclui modelos de regressão de risco proporcional e não proporcional. Desta forma, a partir da formulação na escala de risco, pode-se modelar o risco do evento para o indivíduo i no tempo t adaptando a Equação (4.2) da seguinte forma:

$$h(t|\ddot{\mathbf{x}}_i, \boldsymbol{\zeta_i}) = \sum_{l=1}^{L} \omega_l M_l(t; \boldsymbol{\kappa}, \delta) \exp(\iota + \ddot{\mathbf{x}}_i \boldsymbol{\varphi} + \boldsymbol{\zeta_i} \boldsymbol{\beta}) \ , \tag{4.3}$$

sendo $\boldsymbol{\varrho_2} = (\iota, \boldsymbol{\omega}, \boldsymbol{\varphi}, \boldsymbol{\beta})$ o novo vetor de parâmetros associado ao submodelo estrutural, com $M_l(t; k, \delta)$ denotando o l-ésimo $l = 1, \ldots, L$ termo de base para uma função *M-spline* de grau δ avaliada em um vetor de locais de nó $\boldsymbol{\kappa} = \{\kappa_1, \ldots, \kappa_m\}$, com $m \geq 0$ e ω_l representa o l-ésimo coeficiente *M-spline*. Esta abordagem considera a inclusão de um parâmetro de intercepto no preditor linear (ι) que efetivamente forma parte do risco basal (Brilleman et al., 2020). Para garantir a identificabilidade tanto dos coeficientes *M-spline* quanto do intercepto, que pode ser incluso no preditor linear, os coeficientes *M-spline* são restritos a um simplex, ou seja, $\sum_{l=1}^{L} \omega_l = 1$ (Brilleman et al., 2020). Mais detalhes acerca da definição de *M-splines* são apresentadas no Apêndice C, Seção C.1.

A partir dos desenvolvimentos realizados por Larsen (2004) e Brilleman et al. (2020), é possível reescrever a distribuição de probabilidade de (U_i, ν_i) apresentada na Equação (3.10) como:

$$\begin{aligned} P(u_i, v_i|\ddot{\mathbf{x}}_i, \boldsymbol{\zeta_i}) \propto & \left\{ \sum_{l=1}^{L} \omega_l M_l(u_i; \boldsymbol{\kappa}, \delta) \exp(\iota + \ddot{\mathbf{x}}_i \boldsymbol{\varphi} + \zeta_{\mathbf{i}} \boldsymbol{\beta}) \right\}^{v_i} \\ & \times \exp\left\{ - \sum_{l=1}^{L} \omega_l I_l(u_i; \boldsymbol{\kappa}, \delta) \exp(\iota + \ddot{\mathbf{x}}_i \boldsymbol{\varphi} + \zeta_{\mathbf{i}} \boldsymbol{\beta}) \right\} \ , \end{aligned} \tag{4.4}$$

em que $I_l(u_i; \boldsymbol{\kappa}, \delta)$ denota a l-ésima função de base *I-spline* de grau δ, também denominada *M-spline* integrada, $I_l(u_i; \boldsymbol{\kappa}, \delta) = \int_{\kappa_1}^{u_i} M_l(u_i; \boldsymbol{\kappa}, \delta)\, d\kappa$, que são monotonamente não

decrescentes entre os nós de fronteira (Ramsay, 1988).

Segundo Brilleman et al. (2020), o intercepto do preditor linear (ι) tem como escolhas possíveis as distribuições *a priori* normal, T-Student ou Cauchy, sendo usualmente utilizada uma distribuição $N(0, \sqrt{20})$. No entanto, esta distribuição *a priori* é aplicada ao intercepto após centralizar os preditores em suas médias amostrais, de forma a ajudar na estabilidade numérica e amostragem, mas não impactando nas estimativas reportadas (Brilleman et al., 2020).

Os parâmeros contidos em $\boldsymbol{\beta}$ e $\boldsymbol{\varphi}$ também tem como escolhas possíveis as distribuições *a priori* normal, T-Student ou Cauchy. Neste trabalho, o parâmetro ι e os parâmetros em $\boldsymbol{\beta}$ e $\boldsymbol{\varphi}$ seguem distribuições *a priori* normais, com $\iota \sim N(0, 10^2)$, $\varphi_p \sim N(0, 10^2)$, $p = 1, \ldots, P$ e $\beta_c \sim N(0, 10^2)$, $c = 1, \ldots, C$. A justificativa para adoção dos valores 0 e 10 para os hiperparâmetros relacionados às distribuições *a priori* de ι, e dos parâmeros contidos em $\boldsymbol{\beta}$ e $\boldsymbol{\varphi}$, está relacionada a uma definição utilizando prioris vagas ou não informativas para os efeitos presentes nos vetores $\boldsymbol{\varphi}$ e $\boldsymbol{\beta}$ e no intercepto ι. Já para os parâmetros auxiliares do risco basal via abordagem com *M-Spline* ($\boldsymbol{\omega}$), a opção usual é uma distribuição *a priori* Dirichlet para coeficientes ω_l associados ao risco basal $h_0(t)$ (Brilleman et al., 2020). Desta forma a função de densidade de probabilidade conjunta *a priori* para $\boldsymbol{\omega}$ é dada por:

$$P(\boldsymbol{\omega}) \propto \prod_{l=1}^{L} \omega_l^{\phi_l - 1} \ .$$

4.1.3 Log-verossimilhança para o modelo completo

Considerando-se as definições das subseções anteriores, a nova função de log-verossimilhança para estimação dos parâmetros para o modelo de respostas distais em análise de sobrevivência pode ser expressa através da combinação das metodologias em Brilleman et al. (2020) e Larsen (2004) (Equação (4.4)) e Costa, Amorim e Bispo (2021) (Equação (4.1)), de modo que:

$$
\begin{aligned}
l(\boldsymbol{\varrho}) &= \sum_{i=1}^{N} \left\{ \log \left[P(\mathbf{Y}_i = \mathbf{y}_i, \boldsymbol{\xi}_i = \boldsymbol{\varsigma}_i | \boldsymbol{\gamma}, \boldsymbol{\rho}) \cdot P(u_i, \upsilon_i | \ddot{\mathbf{x}}_i, \boldsymbol{\zeta_i}) \right] \right\} \\
&= \sum_{i=1}^{N} \left\{ \log \left[\prod_{c=1}^{C} \left[\gamma_c \prod_{k=1}^{K} \prod_{r_k=1}^{R_k} \rho_{k,r_k|c}^{I(y_{ik}=r_k)} \right]^{\varsigma_{ic}} \right] + \log \left[P(u_i, \upsilon_i | \ddot{\mathbf{x}}_i, \boldsymbol{\zeta_i}) \right] \right\} \\
&= \sum_{i=1}^{N} \left\{ \sum_{c=1}^{C} \varsigma_{ic} \left[\log(\gamma_c) + \log \left(\prod_{k=1}^{K} \prod_{r_k=1}^{R_k} \rho_{k,r_k|c}^{I(y_{ik}=r_k)} \right) \right] \right. \\
&\quad \left. + \upsilon_i \cdot \left[\log \left(\sum_{l=1}^{L} \omega_l M_l(u_i; \boldsymbol{\kappa}, \delta) \right) + (\eta_i) \right] - \sum_{l=1}^{L} \omega_l I_l(u_i; \boldsymbol{\kappa}, \delta) \cdot \exp(\eta_i) \right\} ,
\end{aligned}
\tag{4.5}
$$

com $\eta_i = \iota + \ddot{\mathbf{x}}_i \boldsymbol{\varphi} + \zeta_{\mathbf{i}} \boldsymbol{\beta}$ representando o preditor linear associado ao submodelo estrutural. Considerando-se o modelo completo, com inclusão dos parâmetros provenientes dos submodelos de mensuração e estrutural, tem-se $\boldsymbol{\varrho} = (\boldsymbol{\rho}, \boldsymbol{\gamma}, \iota, \boldsymbol{\omega}, \boldsymbol{\varphi}, \boldsymbol{\beta})$. Nesta nova log-verossimilhança, apresentada na Equação (4.5), o primeiro termo está associado ao submodelo de mensuração de LCA, enquanto o segundo e terceiro termos estão associados ao modelo semiparamétrico de riscos proporcionais de Cox, com abordagem de *M-splines*. Os métodos propostos no presente trabalho têm como principal objetivo estimar o vetor $\boldsymbol{\beta}$. Mais detalhes do desenvolvimento da Equação (4.5) são apresentados no Apêndice C, Seção C.2.

4.2 Propostas para Estimação dos Parâmetros

Para estimação dos parâmetros do modelo proposto são consideradas duas metodologias: o método Bayesiano Modal Simplificado (BSM) e o método Bayesiano Simultâneo (BS). Ambos consideram o ajuste do modelo de Cox, incluindo uma variável de classe latente (binária ou politômica), além de outras covariáveis, como preditoras do desfecho distal. O método BS apresenta uma vantagem em relação à estimação dos efeitos distais, pois permite a incorporação do erro de mensuração, advindo do modelo de classes latentes, enquanto o método BSM considera as classes latentes preditas modais como observadas, não incorporando o erro de mensuração no modelo estrutural. O objetivo central de ambos os procedimentos é estimar o efeito distal, capturado pelo vetor $\boldsymbol{\beta}$ associado à uma variável latente categórica ζ, preditora não observada do desfecho distal observado, representado pelo tempo até a ocorrência de um evento (T). Além deste vetor de parâmetros, proveniente do submodelo estrutural definido na Subseção 4.1, a metodologia precisa ainda estimar consistentemente os demais parâmetros do vetor $\tilde{\boldsymbol{\varrho}} = (\boldsymbol{\rho}, \boldsymbol{\gamma}, \iota, \boldsymbol{\omega}, \boldsymbol{\varphi})$.

O erro de mensuração não é diretamente incorporado no método BSM, pois considera-se que a variável latente ζ é observada ao se fazer o ajuste do modelo estrutural. As etapas para implementação do método BSM incluem primeiramente a obtenção de uma amostra da distribuição *a posteriori* conjunta dos parâmetros $\boldsymbol{\gamma}$ e $\boldsymbol{\rho}$ da LCA e da variável latente ζ. Daí, pode-se obter as estimativas pontuais para os parâmetros $\boldsymbol{\gamma}$ e $\boldsymbol{\rho}$ através da média da distribuição *a posteriori*. Na próxima etapa os indivíduos são classificados através da moda da distribuição *a posteriori* de ζ, e então estima-se o efeito da classe latente (β) no tempo até ocorrência do evento (resposta distal) por meio do modelo semiparamétrico de Cox, após o cálculo dos *M* e *I-splines*. A estimativa pontual para β é fornecida pela média das amostras da distribuição *a posteriori* do parâmetro, com procedimento análogo sendo realizado para estimativas dos demais parâmetros. Obtém-se ainda o erro padrão para β e demais parâmetros usando as estimativas amostrais obtidas nas iterações. Os passos necessários para a implementação do método BSM são apresentados em detalhes no Algoritmo (3).

Algoritmo 3: Método Simplificado Bayesiano (BSM)

Data: $\{\mathbf{Y}_i(.), \mathbf{U}_i(.), \ddot{\mathbf{X}}_i(.)\}$, $i = 1, \ldots, N$

Result: $\hat{\boldsymbol{\gamma}}$, $\hat{\boldsymbol{\rho}}$, $\hat{\iota}$,$\hat{\boldsymbol{\omega}}$,$\hat{\boldsymbol{\varphi}}$,$\hat{\beta}$

1 **for** *Etapa 1* **do**

2 Ajuste do modelo LCA. A distribuição *a posteriori* dos parâmetros $\boldsymbol{\gamma}$ e $\boldsymbol{\rho}$ e a variável latente $\boldsymbol{\zeta}$ são obtidas utilizando método HMC (via STAN);

3 As estimativas pontuais para $\boldsymbol{\gamma}$ e $\boldsymbol{\rho}$ são definidas como a média de suas distribuições *a posteriori*;

4 Cada indivíduo é classificado na classe latente **que maximiza sua distribuição *a posteriori*** obtida via STAN;

5 **for** *Etapa 2* **do**

6 Definir grau do polinômio δ e os graus de liberdade necessários para cálculo dos *M-splines* e *I-Splines* a partir do tempos observados u_i;

7 Ajuste do modelo semiparamétrico de riscos proporcionais. A distribuição *a posteriori* dos parâmetros ι, $\boldsymbol{\omega}$,$\boldsymbol{\varphi}$ e $\boldsymbol{\beta}$ é obtida utilizando HMC (via STAN);

8 As estimativas pontuais para ι, $\boldsymbol{\omega}$,$\boldsymbol{\varphi}$ e $\boldsymbol{\beta}$ são definidas como a média de suas distribuições *a posteriori*;

9 Os erros padrão para $\hat{\iota}$, $\hat{\boldsymbol{\omega}}$,$\hat{\boldsymbol{\varphi}}$ e $\hat{\boldsymbol{\beta}}$ são definidos como o desvio padrão de suas distribuições *a posteriori*.

O método BS, por sua vez, estima conjuntamente os parâmetros de interesse em um único passo. As etapas para implementação do método BS incluem primeiramente o cálculo dos *M* e *I-splines* através dos tempos observados. Após esta etapa, o modelo semiparamétrico de Cox com abordagem *M-splines* é ajustado aos dados observados ($\mathbf{Y}$,$\mathbf{U}$,$\ddot{\mathbf{X}}$) de forma a maximizar a log-verossimilhança apresentada na Equação (4.5), utilizando

método HMC via *software* STAN, incorporando as distribuições *a priori* escolhidas para os parâmetros. As respectivas estimativas são obtidas através das médias das amostras da distribuição *a posteriori* para os parâmetros do modelo. Desta forma, é possível estimar o efeito da classe latente (β) no tempo até ocorrência do evento (resposta distal) por meio do modelo semiparamétrico de Cox com abordagem *M-splines*, incorporando o erro de mensuração relacionado à variável latente ζ. Obtém-se ainda o erro padrão para β e demais parâmetros usando as estimativas amostrais obtidas nas iterações. Os passos necessários para implementação do método BS são apresentados em detalhes no Algoritmo (4).

Algoritmo 4: Método Bayesiano Simultâneo (BS)

Data: $\{\mathbf{Y}_i(.), \mathbf{U}_i(.), \ddot{\mathbf{X}}_i(.)\}$, $i = 1, \ldots, N$

Result: $\hat{\boldsymbol{\gamma}}$, $\hat{\boldsymbol{\rho}}$, $\hat{\iota}$,$\hat{\boldsymbol{\omega}}$,$\hat{\boldsymbol{\varphi}}$,$\hat{\beta}$

1 **for** *Etapa 1* **do**

2 Definir grau do polinômio δ e os graus de liberdade necessários para cálculo dos *M-splines* e *I-Splines* a partir do tempos observados u_i;

3 Ajuste simultâneo de LCA e modelo semiparamétrico de Cox utilizando método HMC (via STAN);

4 As estimativas pontuais para $\hat{\boldsymbol{\gamma}}$, $\hat{\boldsymbol{\rho}}$, $\hat{\iota}$, $\hat{\boldsymbol{\omega}}$, $\hat{\boldsymbol{\varphi}}$ e $\hat{\beta}$ são definidas como a média de suas distribuições *a posteriori*;

5 Os erros-padrão para $\hat{\boldsymbol{\gamma}}$, $\hat{\boldsymbol{\rho}}$, $\hat{\iota}$, $\hat{\boldsymbol{\omega}}$, $\hat{\boldsymbol{\varphi}}$ e $\hat{\beta}$ são definidos como o desvio padrão de suas distribuições *a posteriori*.

4.3 Implementação em Software Estatístico

A implementação dos *M-splines* e dos *I-splines* (Ramsay, 1988) pode ser realizada através do software R, via pacote `splines2` (W. Wang e Yan, 2024), que, para garantir que a função de risco $h_i(t)$ não seja restrita a zero na origem (ou seja, quando t se aproxima de 0), a base *M-spline* incorpora um intercepto (W. Wang e Yan, 2021, 2024). Além disto, na abordagem para os modelos de sobrevivência na escala de risco apresentada por Brilleman et al. (2020) utiliza-se o grau $\delta = 3$ como padrão, ou seja, *M-splines* cúbicos, de modo que o risco basal possa ser modelado como uma função flexível e suave do tempo. No entanto, vale ressaltar que $\delta = 0$ é um caso especial que corresponde a um risco basal constante por partes.

O processo de implementação em *software* estatístico utilizou ainda os pacotes `RStan` (Stan Development Team, 2024a) e `BRMS` (Bürkner, 2017, 2018, 2021). O `RStan` é uma interface computacional com funções que permitem a compilação, execução e extração de resultados de modelos escritos em linguagem STAN através do *software* R. Já

o `BRMS` é uma interface computacional focada na implementação e sumarização de resultados de modelos, como por exemplo os GLMs, com estimação bayesiana, executados no STAN através do R. Através do `BRMS` foi possível desenvolver implementação inicial do modelos semiparamétrico de Cox com abordagem proposta por Brilleman et al. (2020), e, posteriormente, realizar a integração com os desenvolvimentos de Larsen (2004) e Costa, Amorim e Bispo (2021). O `BRMS` apresenta ainda funções que auxiliam na manipulação de dados, de forma a criar estruturas específicas para bases de dados ou matrizes de acordo com o modelo STAN que se deseje ajustar.

Devido ao tempo extenso de execução dos modelos mediante implementação via STAN foram utilizados também os pacotes `future` e `doFuture` (Bengtsson, 2021) do *software* R. Estes pacotes permitem a execução simultânea ou paralelizada de tarefas mediante a utilização de múltiplos núcleos ou unidades de processamento. Desta forma, foi possível realizar a estimação simultânea e armazenamento de resultados de um número de modelos igual ao número de núcleos de processamento reservados para tal tarefa. Estes pacotes utilizam o conceito de futuros (ou *futures*), que são uma abstração para operações assíncronas que podem ser computadas em paralelo. Eles permitem que se execute tarefas de forma assíncrona e, posteriormente, seja feito o acesso ao resultado quando a computação estiver concluída. Em outras palavras, um *future* é um marcador que representa o resultado de uma operação que ainda está sendo processada e que pode ser acessado no futuro quando o cálculo for finalizado. Mesmo que os modelos não sejam estimados de forma sequencial, os seus resultados são armazenados mantendo a ordenação original.

Para sumarização dos resultados foi também utilizado o pacote `Coda` (Plummer et al., 2006) do *software* R. Este pacote apresenta uma diversidade de funções para análise de resultados e diagnósticos de modelos bayesianos ajustados utilizando abordagens baseadas em MCMC ou HMC. Através das funções deste pacote foram criados gráficos de diagnóstico do modelo, como *traceplots* e curvas de densidade para amostras a posteriori dos parâmetros, gráficos de autocorrelação e gráficos de Gelman-Rubin. Informações acerca dos gráficos de diagnóstico para modelos bayesianos utilizados neste trabalho são apresentadas na Seção 4.4.

4.4 Avaliação de Convergência

Para que as cadeias geradas através de abordagem HMC representem adequadamente as amostras das distribuições a posteriori dos parâmetros de um modelo, elas devem demonstrar propriedades como estacionariedade e convergência, possibilitando inferências válidas e obtenção das estimativas dos respectivos parâmetros. Uma forma de avaliação

das propriedades desejadas nas cadeias geradas por parâmetro é a realização de análise gráfica.

Os gráficos frequentemente utilizados para tal diagnóstico incluem o *traceplot*, gráfico que mostra a evolução dos valores dos parâmetros ao longo das iterações (Brooks e Gelman, 1998); gráfico de densidade da distribuição a *posteriori* dos parâmetros; gráfico de autocorrelação entre as observações amostradas, por cadeia (Plummer et al., 2006); e o gráfico de Gelman-Rubin, relacionado ao número de iterações necessárias para que as cadeias apresentem convergência para um mesmo valor (Gelman e Rubin, 1992; Raftery e Lewis, 1992).

Um *traceplot* é um gráfico que ilustra a evolução dos valores de um parâmetro ao longo das iterações de uma cadeia gerada através dos métodos MCMC ou HMC. Ele é usado para identificar a convergência da cadeia, avaliando se os valores amostrados para um determinado parâmetro se estabilizam em torno de uma média constante após um período inicial. Para que haja indícios de convergência para as diferentes cadeias, com valores iniciais em pontos distintos, deve-se exibir uma oscilação aleatória ao redor de um único valor, sem tendências ou padrões claros. A análise de múltiplas cadeias pode ajudar a verificar se todas convergem para a mesma região, indicando uma boa convergência.

Os gráficos de densidade da distribuição a *posteriori* mostram a distribuição estimada dos parâmetros após a simulação HMC. Esses gráficos permitem visualizar a forma e a dispersão da distribuição a *posteriori*, ajudando a entender a incerteza associada às estimativas dos parâmetros. Para o gráfico de densidade é esperado comportamento unimodal, além de uma distribuição com caudas mais curtas, evidenciando maior certeza nas estimativas obtidas.

Os gráficos de autocorrelação mostram a correlação entre os valores do parâmetro em diferentes iterações da cadeia HMC. Esses gráficos ajudam a identificar a dependência entre amostras sucessivas, uma vez que, ao executar métodos de MCMC ou HMC o valor do parâmetro simulado na iteração $w+1$ depende do valor simulado na w-ésima iteração. Uma baixa autocorrelação sugere que as amostras são independentes e a cadeia está explorando bem o espaço do parâmetro. Alta autocorrelação pode indicar que a cadeia está se movendo lentamente pelo espaço dos parâmetros e pode ser necessário ajustar o algoritmo ou aumentar o número de iterações. O comportamento esperado é que a autocorrelação reduza com aumento do *lag*.

O gráfico de Gelman-Rubin, proposto por Gelman e Rubin (1992), é baseado na estatística $\hat{R}$, conhecida como fator de contração ou redução (Brooks e Gelman, 1998; Raftery e Lewis, 1992). Esta estatística é utilizada para comparar a variabilidade entre as múltiplas cadeias HMC. Seu cálculo é baseado na variação dentro de cada cadeia com a variação entre as cadeias. Valores próximos a 1 indicam que as cadeias convergiram

para a mesma distribuição, enquanto valores significativamente maiores sugerem que as cadeias não convergiram adequadamente. O gráfico ajuda a avaliar a consistência das cadeias, mostrando a partir de quantas iterações as múltiplas cadeias convergiram para um mesmo valor. O comportamento esperado é que a linha do gráfico convirja ao valor 1 à medida que o número de iterações aumenta.

Capítulo 5

Estudos de Simulação

Os métodos bayesianos propostos no Capítulo 4 para estimação do efeito da variável latente categórica na resposta distal do tipo tempo até o evento em dados censurados têm sua performance avaliada através da condução de simulações de Monte Carlo para estudar seu comportamento em amostras finitas. Particularmente, o objetivo dos estudos de simulação foi avaliar o desempenho dos métodos de estimação BSM e BS na estimação do efeito distal β da variável latente categórica ζ no tempo até o evento, utilizando o modelo semiparamétrico de Cox, com uso de *splines* cúbicas para estimação da taxa de falha basal. Apesar da estimação incluir um conjunto maior de parâmetros, o interesse principal é no comportamento do estimador para os β's, que são os coeficientes de regressão associados à variável latente no submodelo estrutural, conforme o modelo de Cox apresentado na Equação (4.3):

$$h(t|\ddot{\mathbf{x}}, \zeta) = \sum_{l=1}^{L} \omega_l M_l(t; \boldsymbol{\kappa}, \delta) \exp(\iota + \ddot{\mathbf{x}}_i \boldsymbol{\varphi} + \boldsymbol{\zeta_i \beta}) \ .$$

A avaliação do estimador β, segundo os métodos bayesianos propostos, é feita considerando-se a definição de diferentes cenários, com variações no tamanho da amostra, no percentual de censura, na magnitude da associação entre a variável latente e o desfecho distal, na entropia do submodelo de mensuração e na presença ou ausência de outras covariáveis no modelo. A entropia é uma característica importante a ser considerada porque está associada ao conceito de homogeneidade e separação das classes latentes. Neste capítulo são explicitados o processo gerador dos dados na Seção 5.1, considerando-se os modelos definidos na Seção 4.1, os cenários para as simulações na Seção 5.2, e os critérios de avaliação na Seção 5.3. Os resultados são apresentados, seguidos da sua discussão, na Seção 5.4.

5.1 Geração dos Dados

Em todos os cenários de simulação, foram gerados 1000 conjuntos de dados com tamanhos amostrais $N = 300$, $N = 500$ e $N = 1000$, utilizando diferentes combinações de valores para os parâmetros dos dois submodelos: mensuração e estrutural. Considerou-se um estudo com tempo total máximo de 104 semanas (unidade de tempo assumida), o que totaliza um período de 2 anos, para a geração dos tempos até o evento. Uma característica importante do submodelo estrutural é o percentual de censura. O mecanismo de censura aleatório à direita foi assumido, pois é o que mais se assemelha a situações encontradas em estudos reais. Os tempos até ocorrência do evento foram gerados segundo abordagem descrita em Bender, Augustin e Blettner (2005), considerando-se uma taxa de falha constante e variando valores de $\boldsymbol{\lambda} \subset [0; 0.25]$, de modo a resultar em percentuais médios de censura de 10 e 30%, respectivamente. Outro aspecto relevante é a presença ou não de variáveis preditoras adicionais à variável latente no submodelo estrutural. Mais detalhes sobre a investigação acerca dos valores de $\boldsymbol{\lambda}$, parâmetro associado aos tempos de falha exponenciais simulados, em relação aos percentuais de censura escolhidos são apresentados no Apêndice (B), Seção (B.1).

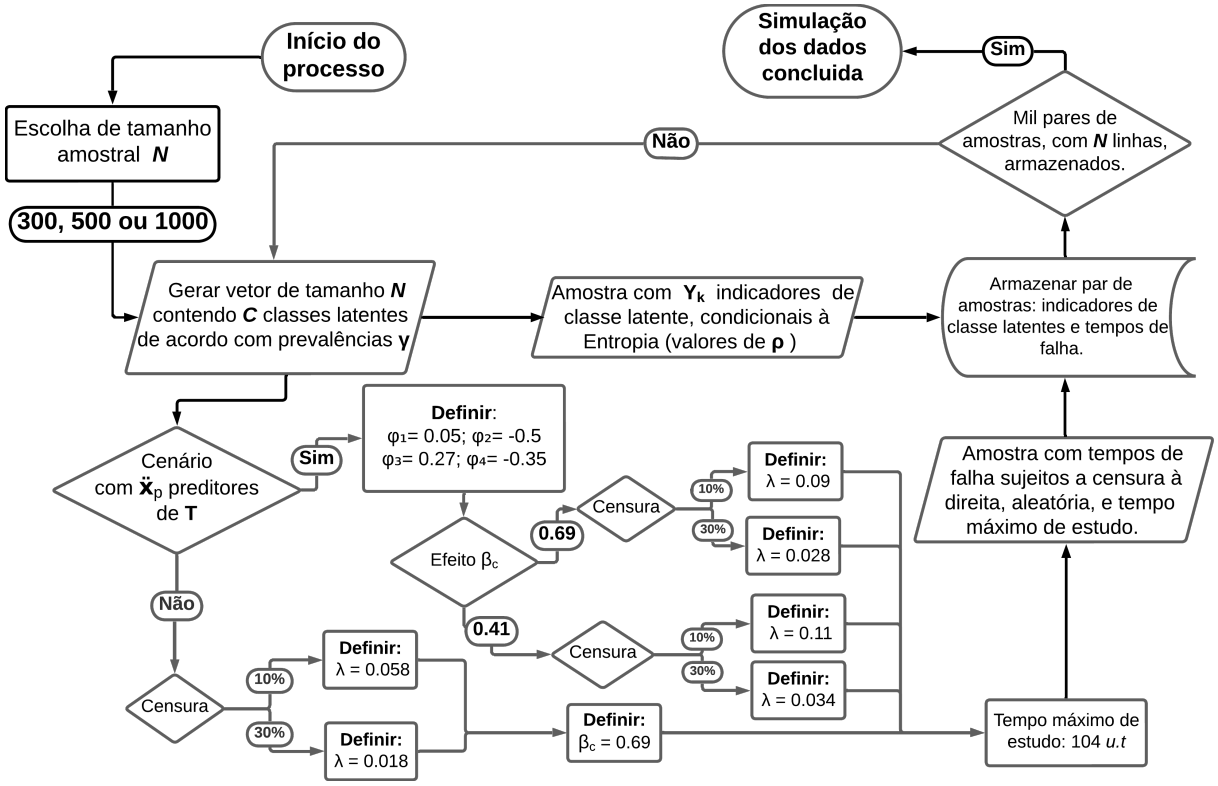
Para a especificação do submodelo de mensuração, foram consideradas 4 variáveis indicadoras binárias, de modo que foram definidos valores para as probabilidades condicionais de resposta ao item ($\boldsymbol{\rho}$) e prevalências de classe ($\boldsymbol{\gamma}$) de acordo com o tipo de entropia média associada, que está diretamente relacionada a uma maior ou menor homogeneidade e separação das classes. Assim, foi investigada uma combinação de parâmetros em que fosse possível obter, através dos modelos de mensuração via LCA, entropias médias padronizadas $\bar{E}$ (Equação (2.34)) iguais a 0.75 (entropia padronizada 'Alta') e 0.45 (entropia padronizada 'Baixa'). Tais valores correspondem à separação de classes alta e baixa, respectivamente (Clark e Muthén, 2009). Desta forma, os cenários com entropia alta e baixa foram definidos como sendo: **(i)** mais homogêneos ($\bar{E}$ em torno de 0.75), com $\boldsymbol{\gamma} = (0.2, 0.8)$, $\boldsymbol{\rho}_1 = (0.8, 0.8, 0.2, 0.2)$ e $\boldsymbol{\rho}_2 = (0.2, 0.2, 0.8, 0.8)$; **(ii)** menos homogêneos ($\bar{E}$ em torno de 0.45) com $\boldsymbol{\gamma} = (0.3, 0.7)$, $\boldsymbol{\rho}_1 = (0.8, 0.6, 0.2, 0.4)$ e $\boldsymbol{\rho}_2 = (0.2, 0.7, 0.6, 0.1)$. Mais detalhes sobre a investigação acerca dos valores de $\boldsymbol{\rho}$ de acordo com a entropia são apresentados no Apêndice (B), Seção (B.2).

Os dados referentes aos indicadores da classe latente e dos tempos de falha foram gerados em duas etapas. Inicialmente, a variável ζ foi especificada através da distribuição multinomial com probabilidades associadas ao vetor de parâmetros $\boldsymbol{\gamma}$, que varia conforme o cenário. A partir de ζ e das probabilidades $\boldsymbol{\rho}$, definidas de acordo com a entropia padronizada média, foram gerados quatro indicadores binários (Y_1 a Y_4). Em seguida, a variável ζ, representando as classes 'verdadeiras' de cada indivíduo, é utilizada como preditora na

simulação dos tempos de falha sujeitos à censura (desfecho distal T), segundo metodologia utilizada em Bender, Augustin e Blettner (2005) para tempos de falha exponenciais.

A Figura 5.1 apresenta o fluxograma do processo de simulação das amostras utilizadas nos cenários apresentados na Tabela 5.1, que é usado até que sejam gerados e armazenados mil conjuntos de dados, cada um contendo pares de amostras (indicadores de classe latente e tempos de falha), para os tamanhos amostrais de 300, 500 e 1000, em cada um dos 12 cenários apresentados. Para os cenários em que não há presença de preditores observados para o desfecho distal, o valor de λ é definido de acordo com o percentual de censura e a magnitude do efeito da variável latente ζ. Para os demais cenários, os valores de λ variam conforme a magnitude do efeito da variável latente ζ e do efeito das demais variáveis preditoras $\ddot{X}_1$, $\ddot{X}_2$ e $\ddot{X}_3$.

Figura 5.1: Fluxograma do processo de simulação das amostras no estudo Monte Carlo.



Fonte: O Autor (2024)

Para facilitar a obtenção da distribuição *a posteriori* dos parâmetros em cenários envolvendo modelos com preditores latentes dicotômicos (Figura 5.2) ou preditores latentes dicotômicos e covariáveis (Figura 5.3), foram utilizadas distribuições *a priori* conjugadas para todos os parâmetros estimados, conforme descrito no Capítulo 4. Para que estas distribuições se tornassem não informativas, foram atribuídos valores iguais a 1 para os parâmetros das distribuições *a priori* Dirichlet, além de médias zero e variâncias 100 para as distribuições normais.

5.2 Cenários e Estimação

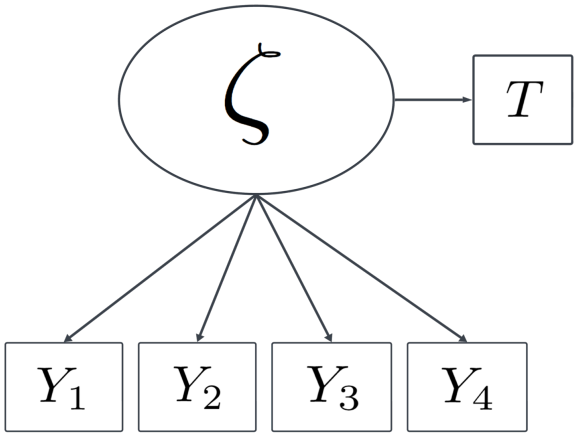
Com o objetivo de investigar a performance do estimador do efeito β da variável latente no desfecho distal do modelo de Cox, usando os métodos propostos BS e BSM, foram definidos 12 cenários, que estão listados na Tabela 5.1. Estes cenários levam em consideração diferentes valores para a magnitude do efeito da variável latente, através do logaritmo da razão das funções taxa de falha ($\beta = 0.69$ ou 0.41), níveis de entropia (Alta/Baixa), percentuais de censura (10% ou 30%) e a presença de preditores (observados) adicionais.

Tabela 5.1: Cenários para os estudos de simulações com duas classes latentes.

$\boldsymbol{\beta}$**: log(HR)**	**Preditores**	**Entropia**	**Censura %**	**Cenários**
0.69	Não	*Alta*	10	***1***
			30	***2***
		Baixa	10	3
			30	4
0.69	Sim	*Alta*	10	***5***
			30	***6***
		Baixa	10	7
			30	8
0.41	Sim	*Alta*	10	***9***
			30	***10***
		Baixa	10	11
			30	12

O modelo teórico apresentado na Figura 5.2 é a base para a geração de dados e o ajuste dos modelos relacionados aos quatro primeiros cenários (1 a 4), que considerou apenas uma variável latente ζ, com $C = 2$ classes, como preditora do desfecho distal.

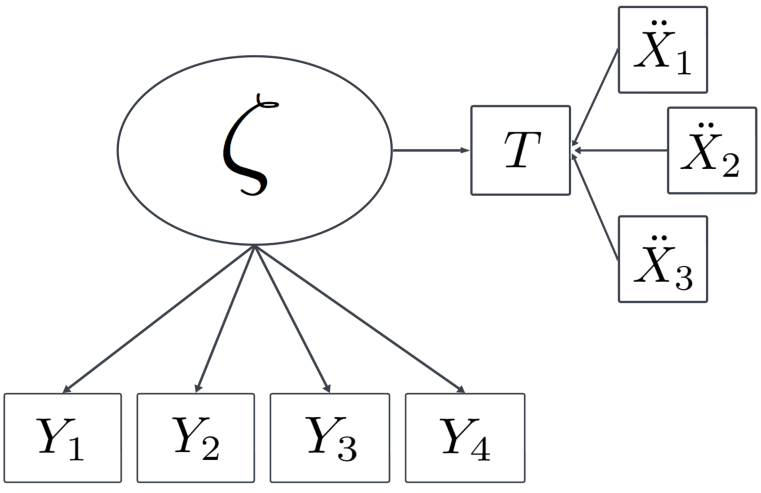
Figura 5.2: Modelo distal considerando LCA com 2 classes latentes sem preditores observados.



Fonte: O Autor (2024)

O modelo conceitual para a geração de dados e o ajuste de modelos que consideram, além da variável latente ζ, com $C = 2$ classes, outros três preditores ($\ddot{\mathbf{X}}_p$, $p = 1, 2, 3$) do desfecho distal é apresentado na Figura 5.3. Para estes cenários (5 a 12) considerou-se $\ddot{X}_1 \sim N(0, 1)$ e $\ddot{X}_2 \sim$ Bernoulli(0.7), com coeficientes, respectivamente, iguais a $\varphi_1 = 0.05$ e $\varphi_2 = -0.5$. A covariável $\ddot{X}_3 \sim$ Multinomial$(0.5, 0.2, 0.3)$ foi incluída no modelo utilizando parametrização que considera o primeiro nível como categoria de referência. Assim, $\varphi_3 = 0.27$ e $\varphi_4 = -0.35$ são os parâmetros associados a $\ddot{X}_3$.

Figura 5.3: Modelo distal considerando LCA com 2 classes latentes e a inclusão de preditores observados.



Fonte: O Autor (2024)

Os valores para os efeitos da variável latente (β =0.41 ou β =0.69) foram escolhidos para representar riscos relativos de, respectivamente, HR=1.50 e HR=1.99, obtidos após a exponenciação de β, denotando, respectivamente, efeitos intermediários e altos (Azuero, 2016). Para a geração dos tempos de falha, foram escolhidos valores de $\boldsymbol{\lambda}$ que resultassem em percentuais médios de censura próximos de 10% e 30%. Esses valores variaram conforme os cenários, sendo $\lambda = 0.058$ e 0.018 para os cenários 1 e 2, e 3 e 4; $\lambda = 0.09$ e 0.028 para os cenários 5 e 6, e 7 e 8; e $\lambda = 0.11$ e 0.034 para os cenários 9 e 10, e 11 e 12, respectivamente.

A estimação do efeito da classe latente β no desfecho distal foi obtida pelos métodos BSM (Bayesiano Modal Simplificado) e BS (Bayesiano Simultâneo), conforme descritos no Capítulo 4. Em síntese, no método BS todos os parâmetros do modelo são estimados de forma simultânea, enquanto no método BSM inicialmente os parâmetros do modelo de mensuração são estimados e, em seguida é extraída a variável de classe predita modal W a partir da maximização a posteriori para a variável latente, levando em conta o número de iterações e cadeias especificadas na função. Daí, em seguida, esta variável é utilizada como se fosse uma variável preditora observada no modelo de sobrevivência.

O STAN Stan Development Team (2024b) foi escolhido como a ferramenta principal para os estudos de simulação devido à sua superior eficiência e flexibilidade na modelagem, especialmente em cenários de maior complexidade, como a estimação simultânea (método BS). Segundo Merkle et al. (2020), a abordagem oferecida pelo STAN proporciona avanços significativos na estimação bayesiana envolvendo variáveis latentes, particularmente em modelos de equações estruturais (SEM), demonstrando ser mais eficiente do que outras alternativas de *softwares* para estimação bayesiana, mesmo em situações desafiadoras envolvendo múltiplas variáveis latentes. Além disso, de acordo com Monnahan, Thorson e Branch (2017), a abordagem do STAN baseada em HMC supera os métodos tradicionais implementados em programas como WinBUGS e JAGS, em termos de eficiência e robustez, especialmente para modelos hierárquicos e de grande complexidade. O STAN não apenas é mais rápido para esses cenários, mas também oferece maior confiabilidade ao detectar problemas de convergência e fornecer diagnósticos detalhados que não estão disponíveis em *softwares* como JAGS ou WinBUGS (Monnahan, Thorson e Branch, 2017). Por essas razões descritas na literatura, o STAN destacou-se como a opção mais adequada para os objetivos deste estudo.

Entretanto, embora o STAN tenha se mostrado a opção mais eficiente para modelos complexos envolvendo variáveis latentes, os estudos de simulação propostos neste trabalho demandam tempo computacional extenso para sua realização devido à quantidade de conjuntos de dados gerados nos diferentes cenários explorados. Assim, para contornar a limitação do tempo de execução dos modelos, foram utilizados os pacotes `future` e

`doFuture` (Bengtsson, 2021), permitindo estimação simultânea de modelos através do uso de computação paralela. Esta abordagem permite a realização simultânea de tarefas que seriam executadas de forma sequencial em um *looping*, mantendo ainda o ordenamento por padrão. Desta forma, foi possível realizar simultaneamente a estimação de 20 modelos (1 modelo por núcleo de CPU) por vez em cada cenário até que fossem estimados os parâmetros e armazenadas as estimativas resultantes dos modelos ajustados com cada uma das 1000 amostras, para os tamanhos amostrais de 300, 500 e 1000. Esta abordagem foi fundamental para obtenção dos resultados em menores intervalos de tempo.

A estimação dos parâmetros dos submodelos de mensuração e estrutural foi realizada utilizando os métodos propostos na abordagem bayesiana, via STAN, com uso das bibliotecas computacionais auxiliares `Rstan` e `BRMS` no *software* R (Bürkner, 2017; Stan Development Team, 2024a), pois as mesmas demonstraram-se úteis durante o processo de estimação dos parâmetros e extração dos resultados dos modelos ajustados.

Mais detalhes acerca dos tempos necessários para estimação, por cenário e tamanho amostral, são apresentados no Apêndice D, Seção D.1. A sintaxe necessária para a estimação em uma e duas etapas de modelos em um mesmo cenário, considerando diferentes tamanhos amostrais, é apresentada no repositório do GitHub `distal_bayes_survival` com estrutura detalhada no Apêndice E.

5.2.1 Aspecto Computacional 1: Rotulação das Classes Latentes

Os estudos de simulação envolvendo modelos de mensuração em LCA compartilham de uma particularidade conhecida na literatura de modelagem com variáveis latentes como '*relabel*' ou '*label switch*' (Redner e Walker, 1984; Stephens, 2000). Isso ocorre porque ao estimar um mesmo modelo várias vezes seguidas pode ocorrer a alternância da ordenação das classes e, consequentemente, daquela em que um indivíduo i é alocado. Para ilustrar considere, como exemplo, um modelo de mensuração com uma variável latente com 3 classes com a implementação do mesmo método de alocação. No primeiro ajuste do modelo um indivíduo i pode ser alocado na classe '1', e no segundo ajuste ser alocado na classe '3', porém tal classe tem as mesmas características de padrão de resposta que a classe '1' resultante da primeira execução do modelo. Logo não há uma troca de grupos e sim do rótulo dos grupos entre diferentes execuções do ajuste do modelo.

Para sanar este problema foram propostas algumas metodologias, sendo uma delas o método da aproximação variacional de Bayes (Ormerod e Wand, 2010). Entretanto, nesta dissertação considerou-se a abordagem de fixação e ordenamento de classes, que visa contornar o problema da troca de rótulos entre classes latentes da seguinte forma: as classes latentes têm seus rótulos reordenados de acordo com suas prevalências, de forma

que a primeira classe tenha a menor prevalência, seguida da segunda, terceira ou demais classes latentes. Desta forma, é possível fixar os rótulos de classe para cada execução do modelo, alterando os mesmos para seguir o padrão pré-estabelecido, mas não alterando o significado e a característica de cada classe latente. Assim, dado um modelo hipotético de LCA com 3 classes latentes, caso um indivíduo i fosse alocado na classe '3' com prevalência de 0.1, enquanto as classes '1' e '2' tem prevalências de 0.3 e 0.6 respectivamente, este indivíduo terá o rótulo da sua classe alterado para '1', e os demais indivíduos alocados inicialmente na classe '1' terão rótulos alterados para '2', e, por fim, os indivíduos restantes que foram alocados inicialmente na classe '2' terão rótulos alterados para '3', preservando a ordenação pré-estabelecida das classes.

Ao realizar a extração dos parâmetros do modelo seguindo a abordagem de ordenação, é importante atentar-se que, ao reordenar os rótulos das classes, é necessário também reordenar o vetor de prevalências das classes latentes e das probabilidades condicionais ao pertencimento às classes. Durante a estimação do efeito da variável latente na resposta distal, caso o modelo estimado apresente a necessidade de '*relabel*' há um procedimento adicional a ser considerado uma vez que o efeito estimado refletirá o logaritmo natural da razão das taxas de falha comparando os indivíduos do grupo de referência (indivíduos na classe com menor prevalência) com os indivíduos na classe de maior prevalência. Para resolução deste problema, o valor do coeficiente do modelo é multiplicado por -1, passando a representar a razão das taxas de falha entre os indivíduos do grupo de maior prevalência tendo como referência os indivíduos classificados no grupo de menor prevalência. Contudo, esta correção é válida apenas para a parametrização usada no modelo de riscos proporcionais de Cox com a variável latente binária.

5.2.2 Aspecto Computacional 2: Convergência

Nestes estudos de simulação considerou-se 5 mil iterações para os cenários com entropia alta, enquanto que nos cenários com entropia baixa foram utilizadas 15 mil iterações. Esses números de iterações foram escolhidos após análises preliminares, que mostraram que nos cenários com baixa entropia é necessário um número mais elevado de iterações para atingir a convergência, principalmente devido à dificuldade em distinguir os indivíduos em classes distintas. Para todos os modelos foi utilizado um *burn-in* equivalente à metade do número de iterações, que é o padrão utilizado pelo STAN. Além disto, como os valores verdadeiros dos parâmetros são conhecidos em cada um dos cenários, foi utilizada 1 cadeia para cada um dos modelos estimados.

Entretanto, antes de utilizar os resultados das estimativas provenientes da distribuição a *posteriori* dos parâmetros é necessário verificar a convergência dos mesmos.

Esta verificação foi realizada neste estudo de simulação através de gráficos como o *traceplot* (adaptado para plotar as estimativas dos parâmetros versus número de amostra) e *boxplots*, de forma a identificar amostras em que a estimativa obtida foi discrepante em relação ao verdadeiro valor do parâmetro e demais estimativas obtidas. Além disto foi realizada verificação da ocorrência de erro de alocação de classe latente, que ocorre quando o modelo aloca todos os indivíduos em uma mesma classe. Desta forma, foram considerados os resultados para 1000 conjuntos de dados em que não ocorreu problema de convergência de parâmetros após estimação dos modelos, sendo os problemas identificados de forma gráfica, via diagnóstico proveniente do pacote `Rstan` ou por verificação de erro de alocação da classe latente. Para atingir esta quantidade de amostras, foi necessário simular amostras complementares em cada cenário. De posse das estimativas adicionais sem problemas de convergência, foi sorteada a quantidade necessária de amostras para substituir as que apresentaram problemas durante a estimação inicial.

Informações adicionais acerca das frequências (absolutas e relativas) dos problemas de convergência detectados ao realizar estimação simultânea dos parâmetros via método BS, por cenário e tamanho amostral, são apresentadas no Apêndice D, Seção D.2.

5.3 Critérios de Avaliação

Vários critérios foram considerados para avaliação das propriedades dos estimadores dos métodos propostos, a citar: estimativa média *a posteriori*, viés relativo percentual, variância da estimativa média *a posteriori*, estimativa média da variância *a posteriori*, raiz do erro quadrático médio e percentual de cobertura. A seguir são apresentadas as definições destas estatísticas e o seu significado para avaliação do efeito da variável latente no desfecho distal.

A estimativa média *a posteriori* denotada por $\hat{\beta}_c$ é obtida pela média das estimativas *a posteriori* do efeito da c-ésima classe latente sobre o desfecho distal, considerando apenas as iterações após *burn-in*, para as mil réplicas em cada tamanho amostral por cenário. Desta forma, seja $r = 1, \ldots, \mathcal{R}$ o índice relacionado ao número de réplicas (modelos estimados), e seja $c = 1, \ldots, C$ a classe latente à qual o efeito distal está relacionado. A estimativa média *a posteriori* $\hat{\beta}_c$ pode ser calculada através da Equação (5.1)

$$\hat{\beta}_c = \frac{\sum_{r=1}^{\mathcal{R}} \widehat{\beta}_{c,r}}{\mathcal{R}} \ . \tag{5.1}$$

O viés relativo percentual, denotado por $VR_{\%}(\hat{\beta}_c)$, foi calculado através da estimativa média *a posteriori* ($\hat{\beta}_c$) para cada tamanho amostral e cenário. O $VR_{\%}(\hat{\beta}_c)$ apresenta as informações sobre o viés de estimação, além do percentual de subestimação ou superestimação das estimativas em cada um dos cenários estudados. Para este critério, valores mais próximos do zero refletem menor viés, valores positivos representam superestimação enquanto valores negativos representam subestimação. Seu cálculo foi realizado através da Equação (5.2):

$$VR_{\%}(\hat{\beta}_c) = \frac{\widehat{\beta_c} - \beta_c}{\beta_c} \times 100 \ . \tag{5.2}$$

A variância do estimador *a posteriori*, denotada por $Var\left(\widehat{\beta}_c\right)$, foi obtida através do cálculo da variância amostral a partir dos valores para as estimativas médias *a posteriori* do efeito distal considerando cada uma das $\mathcal{R}$ réplicas. A Equação (5.3) explicita como foi realizado seu cálculo.

$$Var\left(\widehat{\beta}_c\right) = \frac{\sum_{r=1}^{\mathcal{R}}(\widehat{\beta}_{c,r} - \widehat{\beta}_c)^2}{\mathcal{R} - 1} \ . \tag{5.3}$$

Já a estimativa média da variância *a posteriori*, denotada por $\widehat{Var(\beta_c)}$, é calculada através da média das variâncias estimadas *a posteriori* $\widehat{Var(\beta_{c,r})}$ para todas as $\mathcal{R}$ réplicas. Seu cálculo é feito através do uso da Equação (5.4).

$$\widehat{Var(\beta_c)} = \frac{\sum_{r=1}^{\mathcal{R}} \widehat{Var(\beta_{c,r})}}{\mathcal{R}} \ . \tag{5.4}$$

Para indicar que a variância do estimador está bem estimada, compara-se os valores para os critérios $Var\left(\widehat{\beta}_c\right)$ e $\widehat{Var(\beta_c)}$. Caso apresentem valores semelhantes, esta diferença deve ser próxima a zero.

A raiz do erro quadrático médio, denotado por $RMSE(\widehat{\beta}_c)$, foi calculada levando em consideração os valores das estimativas do efeito $\widehat{\beta}_{c,r}$ em cada uma das $\mathcal{R}$ réplicas. Para este critério de avaliação valores mais próximos de zero são melhores. A Equação (5.5) apresenta sua definição.

$$RMSE(\widehat{\beta}_c) = \sqrt{\frac{\sum_{r=1}^{\mathcal{R}}(\widehat{\beta}_{c,r} - \beta_c)^2}{\mathcal{R}}} \ . \tag{5.5}$$

A probabilidade de cobertura para o efeito β_c, denotada por $CP_{95\%}(\beta_c)$, estima quantas vezes, em $\mathcal{R}$ réplicas, os intervalos de credibilidade associados com as estimativas médias *a posteriori* $\widehat{\beta}_{c,w}$, ao nível de 95%, contém o verdadeiro valor do efeito distal. Esta probabilidade é geralmente expressa em percentual. Para este critério, espera-se que o

$CP_{95\%}(\beta_c)$ seja próximo de 95%. A Equação (5.6) apresenta a formulação necessária para o cálculo da probabilidade de cobertura.

$$CP_{95\%}(\beta_c) = \frac{\sum_{\textit{ż}=1}^{\mathcal{R}} I(\beta_c)_{[LI_{2.5\%}(\hat{\beta}_{c,\textit{ż}});LS_{97.5\%}(\hat{\beta}_{c,\textit{ż}})]}}{\mathcal{R}} \times 100 \ . \tag{5.6}$$

Os termos $LI_{2.5\%}(\hat{\beta}_{c,\textit{ż}})$ e $LS_{97.5\%}(\hat{\beta}_{c,\textit{ż}})$ representam os $\textit{ż}$-ésimos limites inferior e superior, respectivamente, do intervalo de credibilidade associado às $\textit{ż}$-ésimas médias *a posteriori* $\widehat{\beta}_{c,\textit{ż}}$, enquanto $I(\beta_c)$ representa uma função indicadora que assume valor 1 caso β_c esteja contido no intervalo.

5.4 Resultados

As Tabelas 5.2, 5.3 e 5.4 apresentam os resultados dos estudos de simulação relacionados, respectivamente, aos cenários 1 a 4, 5 a 8 e 9 a 12, considerando todos os critérios de avaliação apresentados na Seção 5.3, que são utilizados para avaliar a estimação do efeito distal β_c. De acordo com a parametrização do modelo de riscos proporcionais, o efeito distal representa o logaritmo da razão das funções taxa de falha. Para estimação dos parâmetros do modelo foram utilizados os métodos BSM e BS, considerando os mesmos dados de acordo com o cenário e tamanho amostral.

Tabela 5.2: Estimativas para o coeficiente da variável latente no modelo de Cox em cenários sem a presença de preditores e magnitude alta do efeito ($\beta_c = 0.69$).

Entropia	**Censura**	**N**	***Método BS***						***Método BSM***					
			$\hat{\beta}_c$	$VR_{\%}(\hat{\beta}_c)$	$Var\left(\widehat{\beta_c}\right)$	$\widehat{Var(\beta_c)}$	$RMSE(\widehat{\beta_c})$	$CP_{95\%}(\beta_c)$	$\hat{\beta}_c$	$VR_{\%}(\hat{\beta}_c)$	$Var\left(\widehat{\beta_c}\right)$	$\widehat{Var(\beta_c)}$	$RMSE(\widehat{\beta_c})$	$CP_{95\%}(\beta_c)$
Alta	10%	300	0.689	-0.088	0.035	0.039	0.187	96.0	0.565	-18.064	0.026	0.027	0.204	87.5
		500	0.697	1.079	0.023	0.023	0.152	94.5	0.576	-16.549	0.017	0.016	0.172	83.2
		1000	0.692	0.299	0.010	0.011	0.097	96.6	0.575	-16.727	0.007	0.008	0.144	76.0
	30%	300	0.712	3.225	0.055	0.057	0.236	95.4	0.580	-15.908	0.037	0.037	0.222	90.1
		500	0.700	1.383	0.033	0.032	0.183	95.3	0.579	-16.053	0.022	0.022	0.186	87.1
		1000	0.691	0.194	0.016	0.016	0.126	95.0	0.576	-16.557	0.011	0.011	0.156	80.1
Baixa	10%	300	0.717	3.917	0.082	0.060	0.288	94.2	0.415	-39.876	0.038	0.023	0.338	55.8
		500	0.699	1.274	0.039	0.033	0.198	95.3	0.412	-40.257	0.017	0.012	0.307	31.9
		1000	0.689	-0.204	0.014	0.015	0.116	95.6	0.416	-39.782	0.007	0.006	0.287	6.2
	30%	300	0.760	10.183	0.180	0.475	0.430	94.3	0.421	-38.953	0.035	0.028	0.328	63.2
		500	0.721	4.545	0.056	0.078	0.238	94.3	0.415	-39.872	0.025	0.016	0.317	44.8
		1000	0.705	2.237	0.025	0.021	0.158	93.6	0.424	-38.569	0.011	0.008	0.286	18.4

Analisando as estimativas $\hat{\beta}_c$ no modelo de resposta distal, bem como os respectivos valores de viés relativo percentual ($VR_{\%}(\hat{\beta}_c)$) obtidos pelo método BS, percebe-se que a estimação simultânea dos parâmetros resulta em estimativas de β_c com menor viés em comparação com aquelas obtidas pelo método BSM. Nos dois primeiros cenários, que apresentam alta entropia, observa-se, pelo método BS, uma leve superestimação deste efeito, que diminui à medida que o tamanho amostral aumenta, chegando a ser inferior

a 1% com n=1000. Nestes mesmos cenários, ao considerar o método BSM, há uma subestimação do efeito, o que é um comportamento esperado conforme a literatura a respeito de modelagem com variáveis latentes. No entanto, essa subestimação sofre poucas alterações com o aumento do tamanho amostral, resultando em valores de 16 a 18% menores que o verdadeiro valor de β_c nos melhores cenários.

Nos cenários de baixa entropia, o viés do efeito distal estimado pelo método BS aumenta em amostras menores e percentuais de censura mais elevados, atingindo seu pico com uma superestimação de 10% para um tamanho amostral de 300 e percentual de censura de 30%. Por outro lado, no método BSM, o viés de estimação é mais elevado nos cenários de baixa entropia, reforçando o comportamento já descrito na literatura. Contudo, esse viés varia pouco, mesmo com o aumento do tamanho amostral ou a redução do percentual de censura, levando a estimativas 38 a 40% menores que o verdadeiro valor do efeito.

A variância média da estimador a posteriori, $Var\left(\widehat{\beta_c}\right)$, e a estimativa média da variância a posteriori, $\widehat{Var(\beta_c)}$, apresentam valores semelhantes na maioria dos cenários e métodos de estimação. Ambos os valores, $Var\left(\widehat{\beta_c}\right)$ e $\widehat{Var(\beta_c)}$, diminuem com o aumento do tamanho amostral e são mais elevados em cenários com 30% de censura. Em relação ao $RMSE(\widehat{\beta_c})$, verifica-se uma redução dos valores com o aumento do tamanho amostral, semelhante ao observado para as variâncias estimadas. A $RMSE(\widehat{\beta_c})$ também é maior nos cenários com maior percentual de censura (30%). Por fim, o percentual de cobertura ($CP_{95\%}(\beta_c)$) é maior no método BS em comparação ao BSM. Além disso, à medida que o tamanho da amostra aumenta, o $CP_{95\%}(\beta_c)$ diminui de forma mais acentuada no método BSM, comportamento que não foi observado na estimação via método BS. Essa diminuição observada no $CP_{95\%}(\beta_c)$ está associada a dois fatores: (1) à medida que o tamanho amostral aumenta, a variabilidade das estimativas diminui em ambos os métodos; (2) a redução dessa variabilidade resulta em intervalos de credibilidade com menor amplitude em torno das estimativas $\hat{\beta}_c$, que são de 38 a 40% menores, em cenários com entropia baixa, no método BSM, que não incorpora erro de mensuração. Assim, os intervalos de credibilidade passam a conter menos o verdadeiro valor do parâmetro, pois estão mais concentrados em torno de estimativas até 40% inferiores a esse valor. Este comportamento também é notado no método BSM em cenários com entropia alta, porém de forma menos acentuada uma vez que as estimativas são 14 a 17% menores que o verdadeiro valor do parâmetro.

Tabela 5.3: Estimativas para o coeficiente da variável latente no modelo de Cox em cenários com a presença de três preditores e magnitude alta do efeito ($\beta_c = 0.69$).

Entropia	Censura	N	*Método BS*						*Método BSM*					
			$\hat{\beta}_c$	$VR_{\%}(\hat{\beta}_c)$	$Var\left(\widehat{\beta_c}\right)$	$\widehat{Var(\beta_c)}$	$RMSE(\widehat{\beta}_c)$	$CP_{95\%}(\beta_c)$	$\hat{\beta}_c$	$VR_{\%}(\hat{\beta}_c)$	$Var\left(\widehat{\beta_c}\right)$	$\widehat{Var(\beta_c)}$	$RMSE(\widehat{\beta}_c)$	$CP_{95\%}(\beta_c)$
Alta	10%	300	0.706	2.294	0.038	0.040	0.196	95.1	0.572	-17.030	0.026	0.028	0.200	89.3
		500	0.695	0.796	0.021	0.023	0.146	95.7	0.574	-16.760	0.017	0.017	0.174	83.8
		1000	0.691	0.192	0.012	0.011	0.108	94.4	0.573	-16.946	0.009	0.008	0.151	73.9
	30%	300	0.726	5.195	0.060	0.059	0.248	94.6	0.586	-15.040	0.042	0.039	0.229	89.1
		500	0.714	3.485	0.033	0.033	0.184	95.5	0.590	-14.544	0.024	0.023	0.185	89.2
		1000	0.690	0.064	0.015	0.016	0.122	95.6	0.573	-16.912	0.011	0.011	0.156	81.3
Baixa	10%	300	0.744	7.813	0.080	0.070	0.288	94.7	0.418	-39.400	0.035	0.023	0.330	55.4
		500	0.710	2.957	0.044	0.034	0.211	93.5	0.417	-39.604	0.018	0.013	0.304	33.4
		1000	0.699	1.269	0.015	0.015	0.124	95.1	0.419	-39.293	0.010	0.006	0.290	10.0
	30%	300	0.789	14.297	0.154	0.383	0.404	91.9	0.427	-38.154	0.045	0.030	0.337	65.2
		500	0.735	6.510	0.063	0.078	0.254	94.4	0.418	-39.459	0.026	0.017	0.317	46.2
		1000	0.707	2.405	0.020	0.021	0.144	95.8	0.421	-38.947	0.010	0.008	0.287	17.3

A Tabela 5.3 apresenta os resultados da estimação do efeito da variável latente categórica no tempo até a ocorrência do evento nos cenários em que há presença de preditores observados $\ddot{\mathbf{X}}_p$. Comparando os resultados desta tabela com os da Tabela 5.2, nota-se que, para o método BS, em cenários de alta entropia, o viés do estimador é maior, especialmente para o tamanho amostral de 300. No entanto, esse viés diminui à medida que o tamanho da amostra aumenta. Comparando o viés da estimação pelo método BS com o obtido pelo método em duas etapas, observa-se que o método em uma etapa, apesar de apresentar uma leve superestimação, possui um viés significativamente menor. No método BSM, as estimativas obtidas são entre 15 e 17% menores que o parâmetro. Em relação ao $RMSE(\widehat{\beta}_c)$, o método BS apresenta valores menores do que o método BSM, sendo que, em ambos os métodos, o $RMSE(\widehat{\beta}_c)$ diminui com o aumento do tamanho amostral.

Analisando a Tabela 5.3 nos cenários de baixa entropia, o viés do estimador pelo método BS aumenta em amostras menores e em cenários com maior percentual de censura, atingindo um pico de superestimação de 14% para o tamanho amostral de 300 e 30% de censura, valor 4% superior ao observado na mesma configuração de cenário da Tabela 5.2. Considerando o método BSM, o viés da estimação também é mais elevado em cenários de baixa entropia, comportamento já esperado conforme a literatura e também apresentado na Tabela 5.2. No entanto, assim como em cenários sem preditores observados, esse viés pouco se altera com o aumento do tamanho amostral ou a diminuição do percentual de censura. De acordo com o viés relativo percentual, as estimativas nesses cenários são de 38 a 40% menores que o verdadeiro valor do efeito.

A variância a posteriori do estimador, $Var\left(\widehat{\beta_c}\right)$, e a estimativa média da variância a posteriori, $\widehat{Var(\beta_c)}$, novamente apresentam valores semelhantes em todos os cenários e métodos de estimação. Como mostrado na Tabela 5.3, $Var\left(\widehat{\beta_c}\right)$ e $\widehat{Var(\beta_c)}$ reduzem com o aumento do tamanho amostral e a diminuição do percentual de censura. A probabilidade

de cobertura ($CP_{95\%}(\beta_c)$) é maior no método BS em comparação ao BSM. Além disso, à medida que o tamanho da amostra aumenta, também observa-se uma diminuição mais atenuada no $CP_{95\%}(\beta_c)$ no método BSM, comportamento não observado na estimação via método BS.

Tabela 5.4: Estimativas para o coeficiente da variável latente no modelo de Cox em cenários com a presença de três preditores e magnitude intermediária do efeito ($\beta_c = 0.41$).

Entropia	Censura	N	*Método BS*						*Método BSM*					
			$\hat{\beta}_c$	$VR_{\%}(\hat{\beta}_c)$	$Var\left(\widehat{\beta_c}\right)$	$\widehat{Var(\beta_c)}$	$RMSE(\widehat{\beta}_c)$	$CP_{95\%}(\beta_c)$	$\hat{\beta}_c$	$VR_{\%}(\hat{\beta}_c)$	$Var\left(\widehat{\beta_c}\right)$	$\widehat{Var(\beta_c)}$	$RMSE(\widehat{\beta}_c)$	$CP_{95\%}(\beta_c)$
Alta	10%	300	0.407	-0.835	0.037	0.037	0.193	94.4	0.340	-17.017	0.028	0.027	0.182	92.7
		500	0.414	0.962	0.021	0.022	0.146	94.7	0.346	-15.518	0.016	0.016	0.140	92.6
		1000	0.409	-0.160	0.010	0.010	0.102	95.0	0.342	-16.481	0.008	0.008	0.112	87.6
	30%	300	0.430	4.778	0.057	0.053	0.239	94.7	0.349	-14.780	0.039	0.036	0.206	91.9
		500	0.418	1.919	0.030	0.030	0.173	95.3	0.346	-15.640	0.022	0.021	0.161	92.9
		1000	0.412	0.529	0.014	0.014	0.120	94.9	0.344	-16.151	0.011	0.011	0.122	90.0
Baixa	10%	300	0.443	8.023	0.078	0.068	0.281	93.0	0.263	-35.740	0.026	0.021	0.218	80.9
		500	0.435	6.149	0.038	0.034	0.196	93.9	0.255	-37.775	0.015	0.013	0.197	73.0
		1000	0.418	1.890	0.014	0.013	0.117	94.8	0.253	-38.237	0.007	0.006	0.178	48.5
	30%	300	0.481	17.252	0.131	0.247	0.368	92.4	0.272	-33.542	0.032	0.028	0.226	86.4
		500	0.431	5.114	0.054	0.077	0.232	93.5	0.248	-39.458	0.021	0.016	0.216	73.7
		1000	0.422	2.888	0.022	0.019	0.149	94.5	0.254	-37.974	0.010	0.008	0.184	59.0

A Tabela 5.4 apresenta os resultados da estimação do efeito da variável latente no tempo até a ocorrência do evento nos cenários em que há três preditores observados $\ddot{\mathbf{X}}_p$ e a magnitude do efeito é intermediária. Comparando esses resultados com os da Tabela 5.3, observa-se que, nos cenários de alta entropia, há uma leve subestimação para os tamanhos amostrais de 300 e 1000. Para o tamanho amostral de 500, o comportamento é semelhante ao observado anteriormente, com uma leve superestimação do efeito no cenário com entropia alta e 10% de censura. Em cenários com 30% de censura, o viés de estimação segue o padrão relatado nas Tabelas 5.2 e 5.3. Comparando o viés da estimação pelo método BS com o método em duas etapas, verifica-se novamente que a estimação em uma etapa apresenta um viés menor, com o método BSM produzindo estimativas de 15 a 17% abaixo do verdadeiro valor do parâmetros. Quanto ao $RMSE(\widehat{\beta}_c)$, o método BS continua apresentando valores inferiores aos do método BSM, e, em ambos os métodos, o $RMSE(\widehat{\beta}_c)$ diminui conforme o tamanho amostral aumenta.

Nos cenários de baixa entropia da Tabela 5.4, o viés do estimador pelo método BS cresce em amostras menores e cenários com maior percentual de censura, atingindo 17% de superestimação para o tamanho amostral de 300 e 30% de censura, um aumento de 3% em relação ao cenário correspondente da Tabela 5.3. Pelo método BSM, o viés de estimação é mais elevado, comportamento já previsto na literatura e semelhante ao descrito na Tabela 5.3. Assim como em cenários sem preditores observados, o viés varia pouco com o aumento do tamanho amostral ou a redução do percentual de censura. De acordo

com o viés relativo percentual, as estimativas nesses cenários permanecem entre 38 a 40% abaixo do verdadeiro valor do parâmetro. A avaliação do estimador da variância é similar à já descrita para os demais cenários através da comparação do estimador da variância a posteriori, $Var\left(\widehat{\beta_c}\right)$, com a estimativa média da variância a posteriori, $\widehat{Var(\beta_c)}$. Além disso, como apresentado nas Tabelas 5.2 e 5.3, a probabilidade de cobertura se mantém maior no método BS. No método BSM acréscimos no tamanho amostral também resultam em diminuição mais atenuada no $CP_{95\%}(\beta_c)$.

Capítulo 6

Aplicação

A metodologia proposta no Capítulo 4 pode ser utilizada para responder a diversas questões científicas envolvendo uma variável de exposição latente categórica e um desfecho distal que represente o tempo até o evento em dados censurados. Nesta dissertação ilustra-se o uso destes métodos através da análise de dados provenientes do projeto PrEP1519 para avaliar o efeito do risco real ao HIV (variável latente categórica) no tempo até descontinuidade de uso de profilaxia pré-exposição entre adolescentes homens que fazem sexo com homens (aMSM) e mulheres transexuais (aTGW).

Neste capítulo, são apresentadas informações que contextualizam o tema de pesquisa relacionado ao projeto PrEP1519, bem como as características do estudo descritas na Seção 6.1. As estratégias de análises de dados que são adotadas nesta dissertação são apresentadas na Seção 6.2, enquanto os resultados encontram-se descritos e discutidos na Seção 6.3.

6.1 Contextualização e Dados da Pesquisa

A profilaxia pré-exposição (PrEP) é um método eficaz de prevenção ao Vírus da Imunodeficiência Humana (HIV, em inglês), sendo baseado no uso de uma combinação oral de dose fixa de análogos de nucleosídeos, antirretrovirais Fumarato de Tenofovir Desoproxila (TDF, em inglês) e Emtricitabina (FTC, em inglês), coformulados em um único comprimido (Dourado, Magno et al., 2023). Para o HIV a PrEP é uma opção altamente eficaz para a prevenção (Dourado, Magno et al., 2023; Dourado, Soares et al., 2023), tendo inclusive eficácia demonstrada em ensaios clínicos randomizados (Grant et al., 2010; McCormack et al., 2016).

No final de 2017, o Brasil adotou a PrEP como parte de uma estratégia combinada de prevenção para as populações de maior risco de infecção, nomeadamente homens gays, homens que fazem sexo com homens (HSH), indivíduos transexuais, profissionais do sexo

e usuários de drogas com idade maior ou igual a 18 anos (Dourado, Magno et al., 2023), e, em 2022, a indicação da PrEP foi ampliada para indivíduos da população alvo com idade entre 15 e 17 anos (Dourado, Soares et al., 2023). Contudo, o Brasil não possuía diretrizes específicas para o uso da PrEP entre adolescentes com idade menor a 18 anos e, por isto, pesquisadores de diferentes áreas da saúde conduziram o estudo PrEP1519, o primeiro estudo de coorte de demonstração da PrEP entre adolescentes de população-chave (AKP, em inglês) realizado na América Latina (Dourado, Magno et al., 2023; Dourado, Soares et al., 2023; Zeballos et al., 2023). O estudo foi realizado em três grandes capitais brasileiras: Belo Horizonte (Minas Gerais), Salvador (Bahia) e São Paulo (São Paulo) entre adolescentes homens que fazem sexo com homens e mulheres transexuais, com idades entre 15 e 19 anos. Este estudo tem como objetivo avaliar a eficácia da PrEP em ambientes do mundo real.

Os participantes foram recrutados nas três cidades utilizando diversas estratégias, incluindo o envolvimento de educadores em escolas; idas da equipe até locais onde os jovens se reúnem, uso de aplicativos de conexão e relacionamento, e redes sociais, como Instagram, Facebook, WhatsApp, além do *chatbot* de inteligência artificial chamado "Amanda Selfie", concebida como o primeiro *chatbot* transgênero da América Latina disponível 24 horas por dia, emulando conversas baseadas em bate-papo sobre assuntos delicados, como sexo, doenças sexualmente transmissíveis (DST), PrEP e prevenção combinada (Dourado, Magno et al., 2023).

As estratégias de geração de demanda para recrutamento, inscrição e vinculação a clínicas da PrEP1519 utilizaram uma abordagem de pesquisa-ação para (i) inscrever participantes no estudo e (ii) para analisar a eficácia de diferentes estratégias para recrutar participantes. Os participantes inscritos no estudo tiveram dados coletados digitalmente através de plataforma de registro eletrônico, tornando possível o monitoramento e acompanhamento do participante pela equipe via acesso em tempo real aos dados, garantindo o controle de qualidade das informações coletadas, e respeitando a confidencialidade do participante (Dourado, Magno et al., 2023).

Os participantes compareceram a visitas de acompanhamento após o início do estudo (1º mês e depois a cada três meses). Em cada consulta realizada nas clínicas do estudo foram realizados os procedimentos de avaliação clínica, testes rápidos de HIV, coleta de amostras de sangue para testes sorológicos de infecções sexualmente transmissíveis (IST), como HIV, sífilis, hepatites virais e vírus T-linfotrópico humano (HTLV). Além disto, também foram realizadas avaliações de função renal e hepática para uso de PrEP, adesão cumulativa à terapia baseada em TDF através de coleta de gotas de sangue seco (DBS, em inglês) e esfregaços faríngeos, retais e uretrais para testes de clamídia e gonorréia (Dourado, Magno et al., 2023). Após estes procedimentos, o participante era

encaminhado à unidade de dispensa de medicamento na clínica em que foi realizado o atendimento para a distribuição de PrEP por três meses (3 frascos com 30 comprimidos cada).

Além disto, Dourado, Magno et al. (2023) também descrevem que os participantes com insuficiência renal (taxa de filtração glomerular $< 60mL/min/1,75m^2$, usando a fórmula de Cockcroft-Galt para maiores de 17 anos e a fórmula de Schwartz para menores de 17 anos), história de fratura óssea espontânea, suspeita de síndrome retroviral aguda nos últimos 30 dias, ou relação sexual de risco nas últimas 72 horas foram excluídos temporária ou definitivamente do grupo em uso de PrEP. Nestes casos, foram encaminhados imediatamente para o uso da profilaxia pós-exposição ao HIV (PEP).

Os participantes elegíveis para o uso de PrEP deveriam ter teste anti-HIV negativo e atender a pelo menos um dos seguintes critérios: sexo anal desprotegido nos últimos seis meses, episódio de IST nos últimos 12 meses, uso de PEP nos últimos 12 meses, uso frequente de álcool ou drogas antes ou durante a relação sexual (*chemsex*), troca de dinheiro ou favor por sexo, e qualquer situação específica compartilhada entre adolescente e entrevistador que seja considerada vulnerável ao HIV e outras IST ou experiências relatadas de violência e discriminação devido ao identidade de gênero/orientação sexual. Um questionário sociocomportamental sobre identidade de gênero, acesso a serviços de saúde, práticas sexuais, uso de drogas e álcool e situações de violência foi aplicado no início do estudo e a cada três meses. Os principais desfechos de interesse a serem analisados pelos pesquisadores do estudo são a incidência de HIV, eventos adversos e taxas de adesão e continuação da PrEP. Todos os participantes foram observados até o terceiro ano do estudo (dezembro de 2021) ou até deixarem de participar por opção ou em caso de soroconversão para HIV (Dourado, Magno et al., 2023).

Considerando um dos desfechos principais do projeto PrEP1519, o estudo de Zeballos et al. (2023) teve como objetivo descrever e identificar fatores associados à descontinuação de tratamento preventivo por meio da PrEP como prevenção ao HIV em adolescentes homens que fazem sexo com homens (aMSM) e adolescentes transexuais (aTGW) de 15 a 19 anos. Para a análise realizada em Zeballos et al. (2023) foram considerados dados de 908 adolescentes que iniciaram uso de PrEP, e a descontinuação foi definida como falta de posse de comprimidos de PrEP por período superior a 90 dias. O modelo de regressão de Cox, descrito na seção 3.1.2, foi utilizado para estimar o efeito das covariáveis subpopulação (MSM ou TGW), percepção de risco e ter um parceiro vivendo com HIV no tempo até a descontinuidade no uso de PrEP (Zeballos et al., 2023).

A análise de dados do projeto PrEP1519 nesta dissertação utiliza dados completos disponíveis para 509 participantes que tiveram prescrição de PrEP entre 21 de fevereiro de 2019 e 2 de junho de 2021, e que foram acompanhados até 2 de outubro de 2021. A

descontinuação foi definida como falta de posse de comprimidos de PrEP num período superior a 90 dias, seguindo os critérios definidos por Zeballos et al. (2023). Assim, a variável resposta de interesse é o tempo desde a prescrição até a descontinuação de uso de PrEP. Os participantes foram censurados à direita quando a infecção pelo HIV ocorreu antes da descontinuação ou quando os adolescentes estavam de posse do medicamento PrEP até 2 de outubro de 2021.

Para descrever o perfil de risco real ao HIV (variável latente) dos participantes foram utilizados sete indicadores categóricos binários observados: uso de PEP nos últimos 12 meses, uso de preservativos com parcerias casuais, testagem de HIV com parcerias sexuais, sexo sem penetração, número de parcerias casuais, sexo em grupo e consumo de álcool e outras drogas que interferem no uso de preservativo, todos relacionados aos últimos 3 meses. Com exceção do número de parcerias casuais nos últimos três meses (0:$<$ 4 e 1: $\geq$ 4), as demais variáveis indicadoras foram categorizadas em (0: Não e 1: Sim). Para a estimação do efeito do perfil de risco real ao HIV no modelo para o tempo até a descontinuidade do uso de PrEP, considerou-se cinco covariáveis: idade (0: 15-17 anos e 1: 18-19 anos), escolaridade (0: Ensino superior, 1: Ensino médio e 2: Ensino fundamental), local de estudo (0: Belo Horizonte, 1: Salvador e 2: São Paulo), subpopulação (0: aMSM e 1: aTGW) e raça/cor (0: brancos e 1: não brancos). Foram considerados elegíveis para esta análise todos os participantes que tivessem informações completas sobre as variáveis de linha de base socio-comportamentais, os indicadores de classe latente e as covariáveis.

6.2 Estratégias de Análise

Em uma primeira etapa da análise de dados, a Análise de Classes Latentes (LCA) bayesiana foi implementada para descrever o perfil de risco real ao HIV dos participantes utilizando os sete indicadores binários: uso de PEP nos últimos 12 meses, uso de preservativos com parcerias casuais, testagem de HIV com parcerias sexuais, sexo sem penetração, número de parcerias casuais, sexo em grupo e consumo de álcool e outras drogas que interferem no uso de preservativo, todos relacionados aos últimos 3 meses. Com a finalidade de encontrar o número mais adequado de classes latentes foram ajustados modelos de LCA com 2, 3 e 4 classes, considerando distribuições *a priori* não informativas Dirichilet, definidas conforme Seção 4.1.1, para os parâmetros $\boldsymbol{\gamma}$ e $\boldsymbol{\rho}$ do submodelo de mensuração. Assim como descrito no Capítulo 4, o submodelo estrutural implementado foi o modelo semiparamétrico de Cox usando *M-splines*. Desta forma faz-se necessário definir uma distribuição *a priori* para o vetor $\boldsymbol{\omega}$ de coeficientes associados aos termos base para a função *M-spline*. Nesta análise foi utilizada distribuição *a priori* não informativa

Dirichlet para os coeficientes *M-spline* associados ao risco basal para o modelo de Cox, de acordo com definição apresentada na Seção 4.1.2. A escolha de distribuições *a priori* não informativas foi feita devido ao caráter exploratório desta análise, considerando a ausência de pressupostos prévios sobre a prevalência dos grupos e as probabilidades condicionais. Para identificar o melhor submodelo de mensuração, comparou-se os valores de entropia padronizada (Equação (2.34)), DIC (Deviance Information Criterion), BICM (Bayesian Information Criterion Monte Carlo) e AICM (Akaike Information Criterion Monte Carlo).

Na segunda etapa da análise foi ajustado o modelo para resposta distal para avaliar o efeito dos perfis de risco real ao HIV (variável latente) sobre o tempo até a primeira descontinuidade em uso de PrEP com duas classes latentes. Antes do procedimento de modelagem, os dados do tempo até a primeira descontinuidade foram descritos através da estimação das curvas de sobrevivência pelo método de Kaplan-Meier. Para estimação no modelo com resposta distal foram considerados o Método Bayesiano Modal Simplificado (BSM) e o Método Bayesiano Simultâneo BS, conforme descritos nos Algoritmos 3 e 4 no Capítulo 4. Neste modelo foram incluídas as variáveis preditoras idade, escolaridade, local de estudo, subpopulação e raça/cor. O modelo para o desfecho distal tempo até a descontinuidade foi ajustado conforme descrito no Capítulo 4, através de implementação computacional utilizando sintaxe STAN para uso do HMC na estimação bayesiana, e sintaxe no software R para manipulação e extração dos resultados.

Ao considerar o método BSM, foi estimado modelo de LCA de acordo com o número mais adequado de classes, e através deste modelo os indivíduos foram alocados em classes de acordo com abordagem apresentada no Capítulo 4. Posteriormente essa alocação foi considerada como covariável observada no modelo de riscos proporcionais de Cox. No método BS, por sua vez, a estimação de todos os parâmetros do modelo ocorrem de forma conjunta, incorporando o erro de classificação associado à variável latente, com número de classes igual ao utilizado na estimação segundo método BSM. Para estimação dos referidos modelos também foram utilizados os pacotes RStan (Stan Development Team, 2024a) e BRMS (Bürkner, 2017, 2018, 2021), que são interfaces computacionais que permitem uma maior integração entre a plataforma STAN (Stan Development Team, 2024b) e o software R (versão 4.3.1), fornecendo funções para auxiliar na estimação de modelos, viabilizando extração e análise das suas estimativas.

O ajuste do modelo bayesiano de LCA com resposta distal para os dados do projeto PrEP1519 foi realizado utilizando 10.000 iterações e 8 cadeias. Após o ajuste do modelo foram selecionadas 5 cadeias de forma que as classes latentes com um mesmo significado, tendo como base as probabilidades $\boldsymbol{\rho}$, estejam organizadas de acordo com uma ordenação em comum. Nesta aplicação, as classes latentes referem-se aos perfis de risco real ao HIV. As estimativas dos parâmetros do modelo foram obtidas após verificar se as cadeias

selecionadas para cada um dos parâmetros de interesse apresentam características como estacionariedade e convergência. Para averiguar essas características através das cadeias foram utilizados métodos gráficos como: *traceplot* e curvas de densidade, gráfico de autocorrelação entre observações amostradas e o gráfico de Gelman-Rubin. A avaliação gráfica de convergência foi realizada para os parâmetros do modelo estrutural e para o modelo de mensuração, considerando o método de estimação BS.

6.3 Resultados

Inicialmente foi realizada uma análise descritiva da amostra dos 509 participantes do estudo PrEP1519 considerados nesta dissertação. A Tabela 6.1 apresenta as frequências absolutas e relativas (em percentuais) em cada categoria das variáveis preditoras observadas e indicadores binários das classes latentes. A ordenação das categorias nesta tabela é a mesma utilizada no ajuste do modelo para resposta distal, sendo a primeira categoria o nível de referência para todas as variáveis preditoras observadas.

Tabela 6.1: Descrição da população do Estudo PrEP1519, Brasil. 2019-2021

Características	**Frequência (%)**
Preditores	
Idade	
15-17 anos	98 (19.3%)
18-19 anos	411 (80.7%)
Escolaridade	
Ensino Superior	129 (25.3%)
Ensino Médio	351 (69.0%)
Ensino fundamental	29 (5.7%)
Local de estudo	
Belo Horizonte	126 (24.8%)
Salvador	163 (32.0%)
São Paulo	220 (43.2%)
Subpopulação	
aMSM	464 (91.2%)
aTGW	45 (8.8%)
Raça/cor	
Brancos	132 (25.9%)
Não brancos	377 (74.1%)

Características (Cont.)	Frequência (%)
Indicadores das classes latentes	
Uso de PEP	
Não	452 (88.8%)
Sim	57 (11.2%)
Uso de preservativos com as parcerias casuais nos últimos 3 meses	
Não	44 (8.6%)
Sim	465 (91.4%)
Testagem de HIV com parcerias sexuais	
Não	316 (62.1%)
Sim	193 (37.9%)
Sexo sem penetração	
Não	180 (35.4%)
Sim	329 (64.6%)
Número de parcerias casuais nos últimos 3 meses	
Menor ou igual a 4	304 (59.7%)
Maior que 4	205 (40.3%)
Sexo em grupo nos últimos 3 meses	
Não	371 (72.9%)
Sim	138 (27.1%)
Consumo de álcool e outras drogas de modo a interferir no uso de preservativo	
Não	368 (72.3%)
Sim	141 (27.7%)

Fonte: O Autor (2024)

Analisando as variáveis preditoras, os indivíduos da amostra majoritariamente encontram-se na faixa etária entre 18 e 19 anos (80.7%) e são integrantes da subpopulação aMSM (91.2%). Em relação aos demais preditores, a maioria dos indivíduos tem ensino médio (69%) e é não branca (74.1%) e sua frequência é maior em São Paulo. Dentre as variáveis indicadoras de classe latente, nos últimos 3 meses 44 (8.6%) indivíduos relataram o não uso de preservativos com as parcerias casuais, 205 (40.3%) relataram ter mais que 4 parceiros casuais, e 138 (27.1%) realizaram sexo em grupo. Além disso, 141 (27.7%) relataram interferência no uso de preservativo devido ao consumo de álcool e outras drogas.

A distribuição de frequência dos indicadores das classes latentes evidencia a existência de indivíduos com uso de estratégias de prevenção ao HIV e com comportamentos sexuais

de risco. Dessa forma, a variável latente neste estudo representa perfis de risco real ao HIV. Hipotetiza-se a existência de grupos com maior e menor risco real ao HIV. Assim, o objetivo é avaliar o efeito desses perfis de risco real no tempo até a descontinuidade do uso da PrEP.

Após ajuste do modelo de mensuração de LCA considerando 2, 3 e 4 classes latentes foram computados os valores para estatísticas de ajuste DIC, BICM, AICM além da entropia padronizada (Tabela 6.2). Com base nas estatísticas de ajuste apresentadas na Tabela 6.2, o modelo com 2 classes latentes foi escolhido por apresentar melhores valores para o DIC e BICM além de entropia de 0.4610. Os modelos com 3 e 4 classes latentes, apesar de terem melhores AICM, ou seja, valores menores, apresentam entropias inaceitavelmente baixas, com valores de 0.0825 e 0.0032, respectivamente.

Tabela 6.2: Estatísticas de ajuste para modelos bayesianos de LCA para os perfis latentes de risco real ao HIV. Projeto PrEP. 2019-2021.

Estatísticas	**Número de classes**		
	2	**3**	**4**
Entropia	0.4610	0.0825	0.0032
AICM	-3811.231	-3815.504	-3815.303
BICM	-3873.834	-3951.029	-3957.503
DIC	-3810.525	-3755.34	-3789.874

Na Tabela 6.3 são apresentadas as estimativas do submodelo de mensuração - as prevalências das classes latentes e as probabilidades de resposta ao item - obtidas usando o método BS com duas classes latentes. Nesta análise para os dados do projeto PrEP1519 as variáveis indicadoras são relacionadas a estratégias de prevenção ao HIV e a comportamentos sexuais de risco, entretanto, probabilidades mais elevadas para os indicadores relacionados ao uso de preservativos com parcerias casuais nos últimos 3 meses e sexo sem penetração refletem um menor risco real ao HIV, enquanto que probabilidades condicionais mais elevadas para uso de PEP (utilizada após qualquer situação em que exista risco de contágio de HIV) nos últimos 12 meses, testagem de HIV com parcerias sexuais, número de parcerias casuais e sexo em grupo nos últimos 3 meses, respectivamente, e consumo de álcool e outras drogas que interferem no uso de preservativo, estão relacionadas com um maior risco real ao HIV. Desta forma, as probabilidades condicionais de resposta ao item podem ser utilizadas como base para distinção e caraterização das classes em dois perfis de risco real ao HIV, aqui denominados como comportamento de baixo risco (classe 1) ou alto risco (classe 2).

Tabela 6.3: Estimativas do submodelo de mensuração da LCA com resposta distal, em uma etapa, com abordagem bayesiana para o risco real ao HIV. Projeto PrEP. 2019-2021.

Risco real ao HIV	**Prevalências (γ)** [95% ICr]	
	Baixo: 0.42 [0.30 ; 0.56]	**Alto**: 0.58 [0.44 ; 0.70]
Indicadores das classes latentes	**Prob. Condicionais ($\rho_{k,c}$)** [95% ICr]	
Uso de PEP		
Sim	0.08 [0.04 ; 0.12]	0.14 [0.10 ; 0.18]
Uso de preservativos com as parcerias casuais nos últimos 3 meses		
Sim	0.83 [0.78 ; 0.88]	0.96 [0.94 ; 0.99]
Testagem de HIV com parcerias sexuais		
Sim	0.24 [0.14 ; 0.33]	0.47 [0.42 ; 0.53]
Sexo sem penetração		
Sim	0.47 [0.37 ; 0.56]	0.77 [0.71 ; 0.82]
Número de parcerias casuais nos últimos 3 meses		
Maior que 4	0.09 [0.02 ; 0.17]	0.64 [0.54 ; 0.76]
Sexo em grupo nos últimos 3 meses		
Sim	0.06 [0.01 ; 0.12]	0.43 [0.36 ; 0.51]
Consumo de álcool e outras drogas de modo a interferir no uso de preservativo		
Sim	0.19 [0.13 ; 0.25]	0.34 [0.29 ; 0.40]

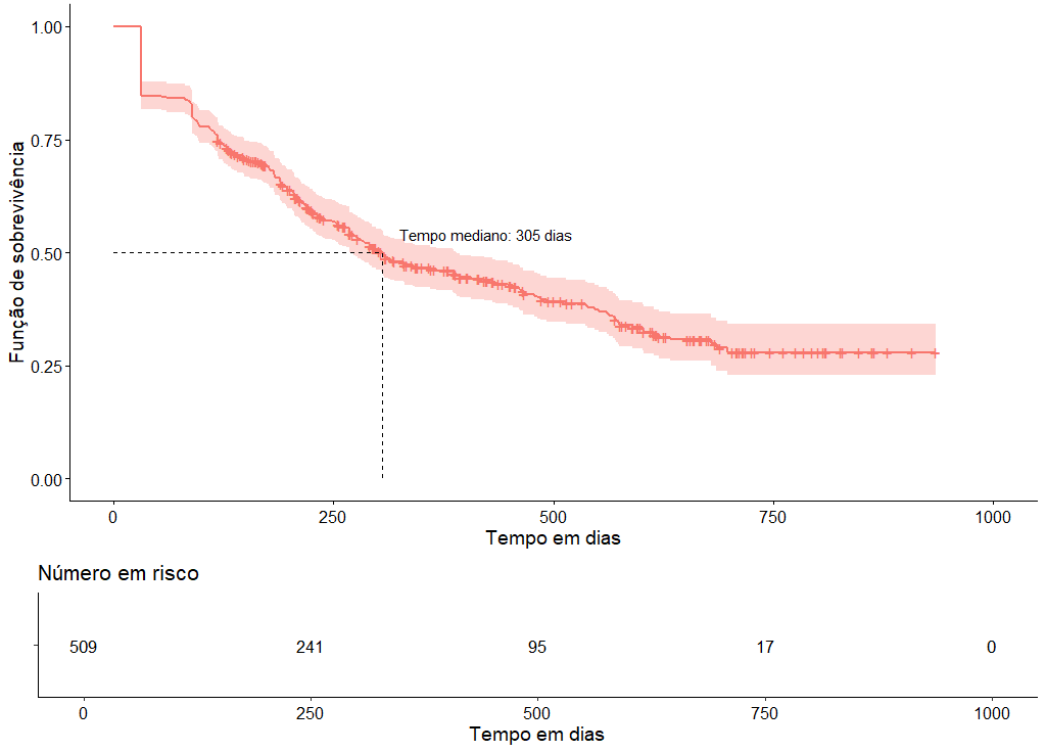
De acordo com a Tabela 6.3, as probabilidades de uso de PEP dado pertencimento aos perfis de risco real ao HIV, denominados de baixo e alto risco, foram de 0.08 e 0.14, respectivamente. Em relação ao uso de preservativos com parceiros casuais nos últimos 3 meses, tem-se probabilidade elevada em ambos os perfis, contudo, o perfil de maior risco real ao HIV apresentou probabilidade um pouco mais elevada se comparado com o perfil de baixo risco, e isto pode estar relacionado ao fato de que 91.4% dos indivíduos da amostra relatam o uso de preservativo com parcerias casuais, sendo um pouco maior no grupo de maior risco (96%). Comportamento semelhante é observado em relação à prática de sexo sem penetração, em que o perfil de maior risco apresentou maior probabilidade condicional (77%), enquanto esta prática foi relatada por (47%) dos indivíduos caracterizados como tendo menor risco real ao HIV.

Em relação às probabilidades condicionais associadas aos demais indicadores apresentados na Tabela 6.3, a probabilidade de um indivíduo realizar testagem de HIV com parcerias sexuais é de 47% no perfil de maior risco e de 24% no perfil de menor risco. A probabilidade de se ter mais que 4 parcerias casuais nos últimos 3 meses é de 64% para o perfil de maior risco, enquanto para o perfil de menor risco essa probabilidade é de 9%. Em relação a ter participado de sexo em grupo nos últimos 3 meses, as probabilidades condicionais de resposta positiva para os perfis de risco alto e baixo foram de, respecti-

vamente, 43% e 6%. Por fim, as probabilidades condicionais de resposta positiva para interferência de álcool e outras drogas no uso de preservativo para os perfis de risco alto e baixo foram de, respectivamente, 34% e 19%.

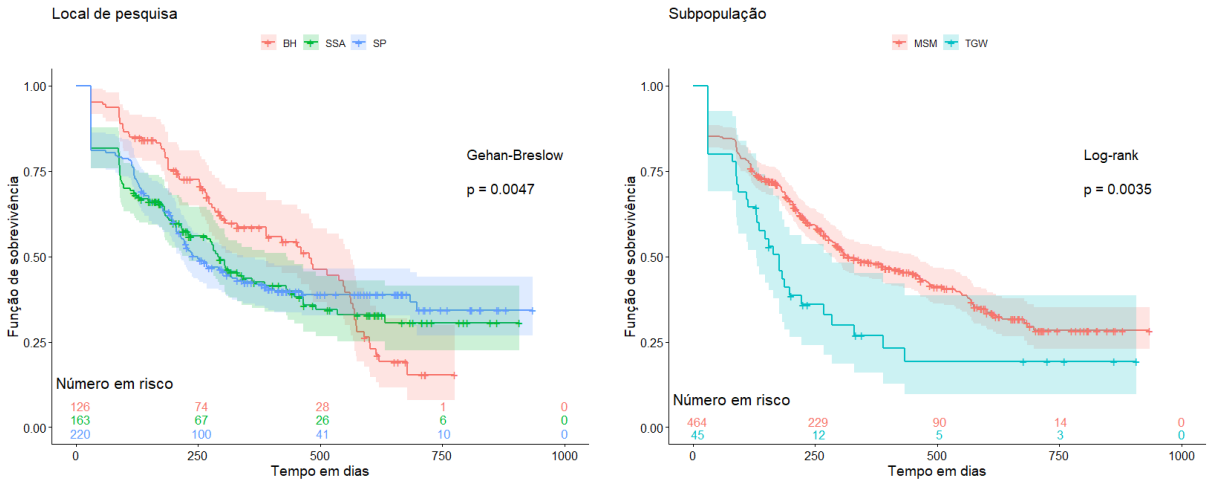
Antes do ajuste do modelo para a resposta distal, foram realizadas algumas análises descritivas relacionadas ao tempo até a primeira descontinuidade da PrEP. A Figura 6.1 apresenta o gráfico contendo as probabilidades estimadas globais de sobrevivência, enquanto as Figuras 6.2 a 6.3 apresentam as estimativas da função de sobrevivência pelo método de Kaplan-Meier, segundo as variáveis preditoras.

Figura 6.1: Função de sobrevivência estimada para o tempo até a descontinuidade da PrEP usando método de Kaplan-Meier. Projeto PrEP1519. 2019-2021.



A Figura 6.1 apresenta a função de sobrevivência estimada, que descreve a probabilidade de não descontinuidade da PrEP durante os meses do estudo. É possível perceber o decaimento no número de indivíduos sob risco em cada ponto do tempo, seja por terem descontinuado ou por censura. A linha pontilhada em preto está relacionada com o tempo mediano de sobrevivência. Assim, verifica-se que 50% dos indivíduos não descontinuaram em até 305 dias, ou seja, antes do primeiro ano de acompanhamento. Após um ano de estudo apenas 155 (30.5%) dos indivíduos ainda estavam sob risco de descontinuidade da PrEP. O percentual de censura foi de 42.2%.

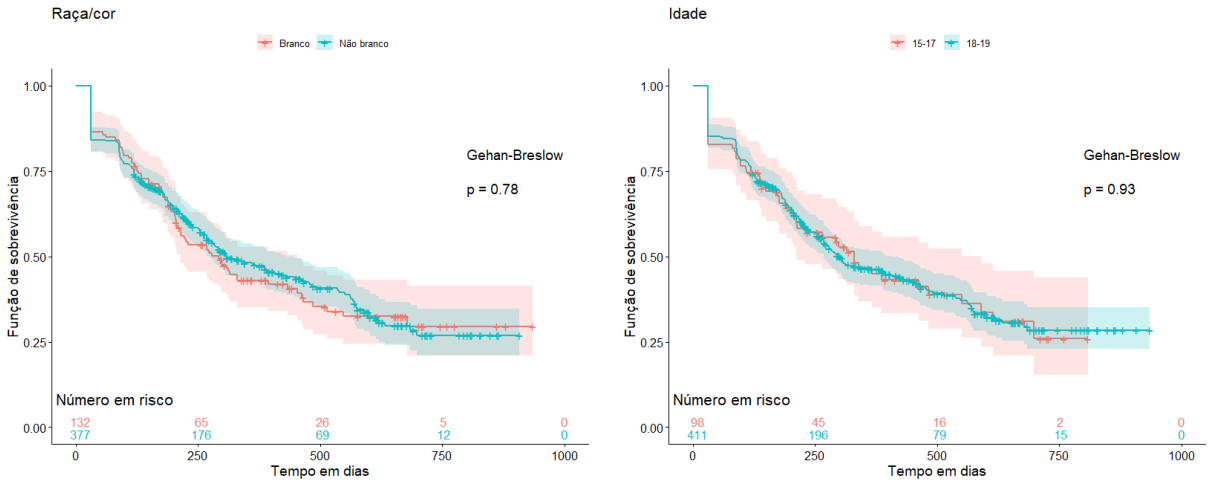
Figura 6.2: Função de sobrevivência estimada para o tempo até a descontinuidade da PrEP, segundo subpopulação e local do estudo. Projeto PrEP1519. 2019-2021.



(a) (b)

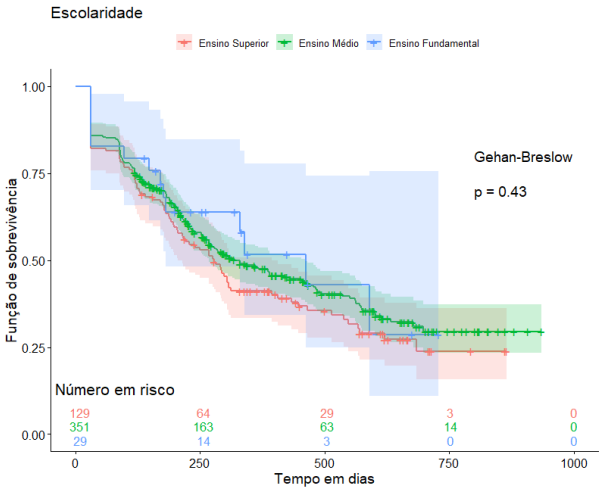
Analisando as curvas de sobrevida por local do estudo, verifica-se na Figura 6.2a que Belo Horizonte destaca-se pelas maiores probabilidades de sobrevivência, ou seja, de não descontinuidade até por volta do décimo oitavo mês. Após este período, o reduzido número de participantes em cada local inviabiliza conclusões relevantes. Já as análises segundo a subpopulação, apresentadas na Figura 6.2b, evidenciam que os adolescentes transgênero têm uma probabilidade de continuidade menor que os adolescentes aMSM em todos os pontos do tempo. Aplicando-se o teste de Gehan-Breslow e Log-rank, respectivamente, para comparar as funções de sobrevivência segundo o local e a subpopulação, ao nível de 5% de significância, foram encontradas evidências de diferença estatisticamente significativa entre estes grupos.

Figura 6.3: Função de sobrevivência estimada para o tempo até a descontinuidade da PrEP, segundo raça, idade e escolaridade. Projeto PrEP1519. 2019-2021.



(a)

(b)



(c)

Analisando as curvas de sobrevida por raça/cor e idade, respectivamente, nas Figuras 6.3a e 6.3b, é possível perceber sobreposição entre as curvas ao longo do tempo, indicando a inexistência de diferença nos tempos de descontinuidade entre as categorias de raça/cor e faixas etárias. Estes resultados são corroborados pelo teste de Gehan-Breslow, que, em ambos os casos, não detectou diferença estatisticamente significante, ao nível de 5% de significância, entre as funções de sobrevivência, ou seja, as probabilidades de não descontinuidade de tais grupos. As curvas de sobrevida para escolaridade, apresentadas na Figura 6.3c, indicam comportamento com menor sobreposição. No entanto, não há evidências de diferença estatisticamente significante na função de sobrevivência por nível de escolaridade segundo o resultado do teste de Gehan-Breslow, ao nível de 5% de significância. Ressalta-se que apenas para as análises por subpopulação considerou-se o

teste de Log-rank devido ao seu pressuposto de proporcionalidade dos riscos, que não foi atendido para as outras variáveis.

Na Tabela 6.4 são apresentadas as estimativas do submodelo estrutural bayesiano em uma etapa com resposta distal com 2 classes latentes. Os coeficientes do modelo de Cox, com inclusão dos intervalos de 95% de credibilidade, encontram-se na Tabela 6.4, através dos quais é possível avaliar a magnitude dos efeitos de diminuição (estimativas menores que 1) ou acréscimo (estimativas maiores que 1) para o risco de descontinuidade do tratamento preventivo para HIV, com uso de PrEP. O coeficiente associado à razão das taxas de falha do perfil de maior risco real ao HIV, em comparação ao de menor risco, evidencia que os indivíduos alto risco real ao HIV têm 41% menos risco de descontinuar ou abandonar o tratamento com uso de PrEP se comparados com os indivíduos de baixo risco real $IC_{95\%}(HR) = [0.42; 0.81]$. Esse comportamento, embora possa parecer contraintuitivo, já que os indivíduos com comportamento de maior risco descontinuam menos, está alinhado com a literatura da área. Uma revisão sistemática recente (Zhang et al., 2022) identificou, em diversos estudos, que um menor risco clínico de infecção por HIV é uma razão comum para a interrupção da PrEP (Zeballos et al., 2023).

Analisando os demais coeficientes do modelo estrutural apresentados na Tabela 6.4, verifica-se que indivíduos com idade entre 18 e 19 anos apresentaram 3% menos risco de descontinuação do tratamento se comparados com os indivíduos com idade entre 15 e 17 anos $IC_{95\%}(HR) = [0.74; 1.27]$. Como já relatado na literatura, menores idades estão frequentemente relacionados à interrupção da PrEP (Zeballos et al., 2023; Zhang et al., 2022). Segundo a escolaridade, os indivíduos com ensino médio e ensino fundamental apresentaram reduções de 19 e 36%, respectivamente, no risco de descontinuidade do tratamento se comparados com os indivíduos com escolaridade de nível superior. Para o local do estudo, os indivíduos que realizam o tratamento nas cidades de Salvador e São Paulo apresentaram riscos 23 e 20% maiores, respectivamente, de descontinuidade do tratamento se comparados com indivíduos que realizam o tratamento na cidade de Belo Horizonte. Em relação a subpopulação (gênero), os adolescentes aTGW apresentam risco 70% maior de descontinuação se comparados com adolescestes aMSM $IC_{95\%}(HR) = [1.23; 2.34]$. Ao analisar a raça/cor autodeclarada, os indivíduos não brancos apresentaram um risco de descontinuação 7% menor se comparados com os autodeclarados brancos.

Tabela 6.4: Estimativas do modelo de LCA com respostas distais, em uma etapa, com abordagem bayesiana, para avaliar o efeito de fatores de risco na descontinuidade da PrEP. 2019-2021.

Variáveis preditoras	**HR**[a]	**[95% CrI]**
Risco real ao HIV (Latente)		
Baixo	Ref	
Alto	0.59	[0.42 ; 0.81]
Idade		
15-17 anos	Ref	
18-19 anos	0.97	[0.74 ; 1.27]
Escolaridade		
Ensino Superior	Ref	
Ensino Médio	0.81	[0.65 ; 1.02]
Ensino fundamental	0.64	[0.38 ; 1.05]
Local de estudo		
Belo Horizonte	Ref	
Salvador	1.23	[0.94 ; 1.62]
São Paulo	1.20	[0.93 ; 1.55]
Subpopulação		
aMSM	Ref	
aTGW	1.70	[1.23 ; 2.34]
Raça/cor		
Brancos	Ref	
Não brancos	0.93	[0.73 ; 1.17]

Para fins de comparação, a análise dos dados do projeto PrEP1519 também foi realizada considerando o método de estimação BSM (em duas etapas), e os resultados do submodelo estrutural estão apresentados na Tabela 6.5. Não houve diferenças relevantes nas estimativas dos parâmetros do submodelo de mensuração (LCA), sendo mantidas as mesmas interpretações realizadas mediante estimação via método BS (em uma etapa). Para os parâmetros do submodelo estrutural, que estão relacionados às razões das funções de taxa de falha dos preditores observados, contudo, verifica-se a subestimação dos efeitos, como esperado mediante literatura de variáveis latentes e os estudos de simulação realizados no Capítulo 5. Assim, houve uma subestimação do efeito associado ao preditor latente (risco real ao HIV), com uma redução do risco de descontinuidade para os indivíduos com alto risco, que passou de 41% (considerando estimação simultânea via método BS) para 19% (considerando estimação em duas etapas via método BSM).

Tabela 6.5: Estimativas do modelo de LCA com respostas distais, em duas etapas, com abordagem bayesiana, para avaliar o efeito de fatores de risco na descontinuidade da PrEP. 2019-2021.

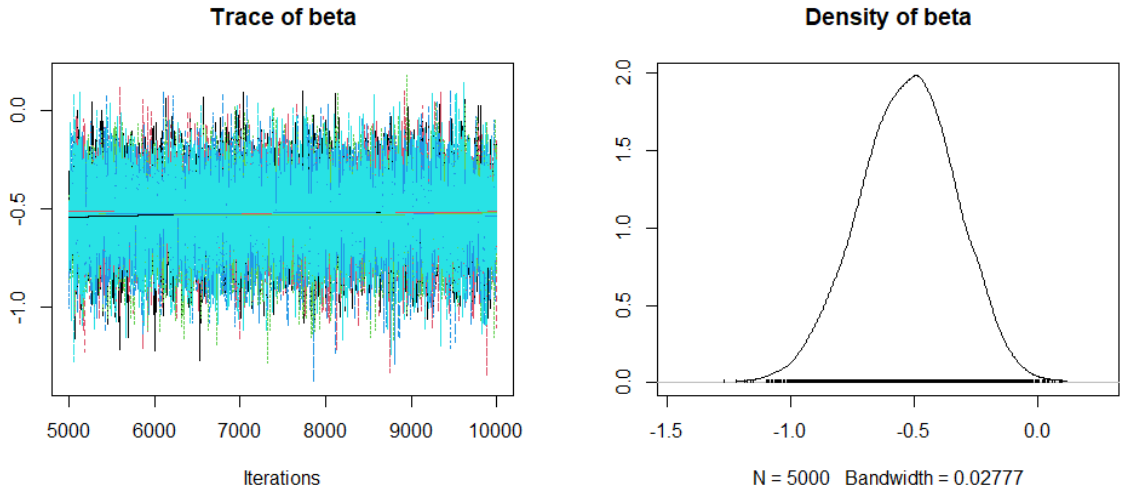
Variáveis preditoras	**HR**[a]	**[95% ICr]**
Risco real ao HIV (Latente)		
Baixo	Ref	
Alto	0.81	[0.64 ; 1.02]
Idade		
15-17 anos	Ref	
18-19 anos	0.98	[0.71 ; 1.35]
Escolaridade		
Ensino Superior	Ref	
Ensino Médio	0.83	[0.64 ; 1.09]
Ensino fundamental	0.68	[0.36 ; 1.21]
Local de estudo		
Belo Horizonte	Ref	
Salvador	1.22	[0.89 ; 1.67]
São Paulo	1.16	[0.87 ; 1.57]
Subpopulação		
MSM	Ref	
TGW	1.66	[1.13 ; 2.38]
Raça/cor		
Brancos	Ref	
Não brancos	0.95	[0.73 ; 1.25]

Risco real ao HIV	**Prevalências (γ)** [95% ICr]	
	Baixo: 0.47 [0.34 ; 0.62]	**Alto**: 0.53 [0.38 ; 0.66]
Indicadores de classe latente	**Prob. Condicionais ($\rho_{k,c}$)** [95% ICr]	
Uso de PEP		
Sim	0.08 [0.05 ; 0.12]	0.15 [0.11 ; 0.19]
Uso de preservativos com as parcerias casuais nos últimos 3 meses		
Sim	0.85 [0.79 ; 0.89]	0.97 [0.94 ; 0.99]
Testagem de HIV com parcerias sexuais		
Sim	0.27 [0.18 ; 0.35]	0.47 [0.41 ; 0.54]
Sexo sem penetração		
Sim	0.50 [0.40 ; 0.58]	0.78 [0.72 ; 0.84]
Número de parcerias casuais nos últimos 3 meses		
Maior que 4	0.11 [0.03 ; 0.19]	0.68 [0.56 ; 0.81]
Sexo em grupo nos últimos 3 meses		
Sim	0.07 [0.01 ; 0.14]	0.46 [0.38 ; 0.55]
Consumo de álcool e outras drogas de modo a interferir no uso de preservativo		
Sim	0.20 [0.14 ; 0.26]	0.35 [0.3 ; 0.41]

A seguir são apresentados gráficos para verificação de ajuste e convergência para o submodelo estrutural aplicado à análise dos dados do projeto PrEP1519 segundo método BS de estimação. Através destes gráficos é possível avaliar se as distribuições a posteriori simuladas convergiram para uma distribuição estacionária, e para isto foram utilizadas cinco cadeias com valores iniciais distintos para que seja possível observar se estas estão convergindo para um mesmo valor.

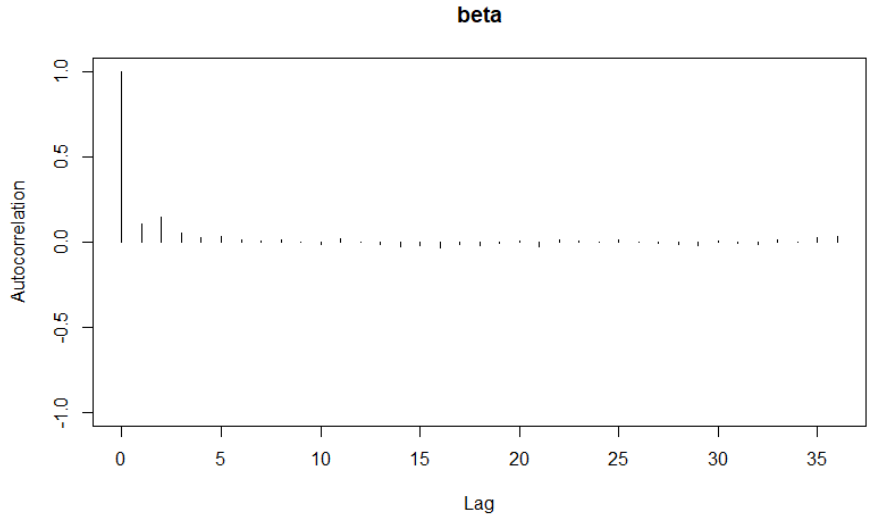
As Figuras 6.4, 6.5 e 6.6 apresentam os gráficos de trajetória (*traceplot*) e densidade, autocorrelação e Gelman-Rubin, respectivamente, relacionados ao parâmetro β, associado ao preditor latente ζ que representa os perfis de risco real ao HIV, segundo método de estimação BS, com resultados apresentados na Tabela 6.4.

Figura 6.4: *Traceplot* e curva de densidade para parâmetro β, de acordo com 5 cadeias geradas pelo método BS para os dados PrEP. 2019-2021



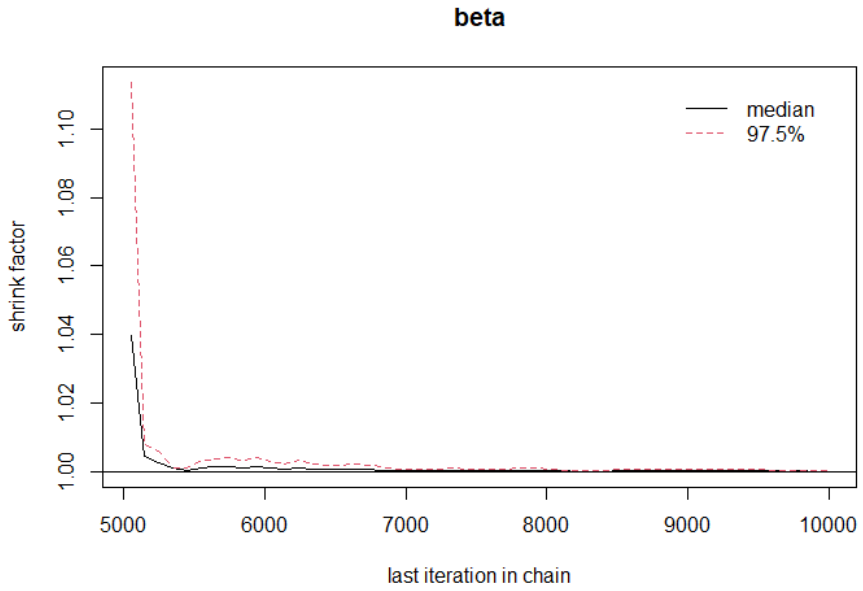
Analisando a Figura 6.4 que apresenta o *traceplot* e a curva de densidade associados ao parâmetro β, percebe-se que as cadeias apresentadas misturam-se adequadamente, não sendo identificadas tendências entre as cadeias exploradas e apresenta valores oscilando em torno de uma média estável. Além disto, o gráfico de densidade da distribuição *a posteriori* para β apresenta apenas uma moda, e está distribuído simetricamente em torno de um valor, apresentando caudas curtas, o que reflete numa maior assertividade no cálculo da estimativa.

Figura 6.5: Autocorrelação para valores amostrados do parâmetro β, de acordo com uma cadeia, gerada pelo método BS para os dados PrEP. 2019-2021



Analisando a Figura 6.5 nota-se que a partir do *lag* 5 não são detectadas correlações expressivas entre as amostras da distribuição *a posteriori* para β, o que é um indicativo de que a autocorrelação diminui rapidamente com o aumento do *lag*. Desta forma tem-se o indicativo de amostras da distribuição *a posteriori* independentes. Logo, uma cadeia com menos iterações, mas com valores independentes, consegue fornecer informações suficientes para cálculo da estimativa para o parâmetro β. O mesmo tipo de gráfico para as demais cadeias apresenta comportamento análogo ao gráfico apresentado com a primeira cadeia.

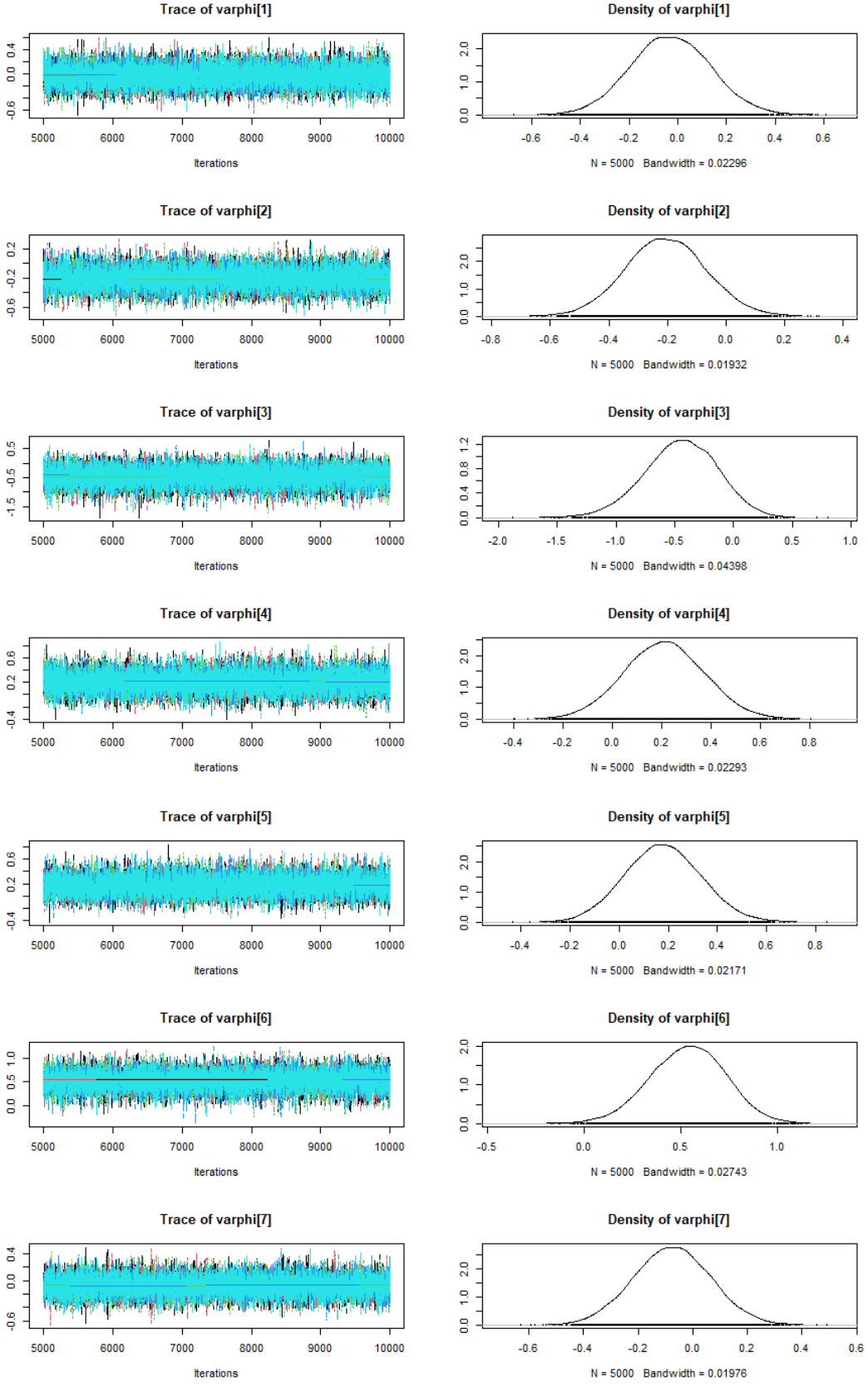
Figura 6.6: Gráfico de estatística Gelman-Rubin para o parâmetro β, de acordo com 5 cadeias geradas pelo método BS para os dados PrEP. 2019-2021



A Figura 6.6 apresenta o diagnóstico de Gelman-Rubin, representado pelo valor $\hat{R}$ comparando a variabilidade entre múltiplas cadeias HMC com a variabilidade dentro de cada cadeia, avaliando se as diferentes cadeias convergiram para a mesma distribuição. A partir do gráfico tem-se evidências de que as cadeias convergem para a mesma distribuição posterior após 5000 iterações, uma vez que o fator de redução de escala em potencial se aproxima do valor 1.

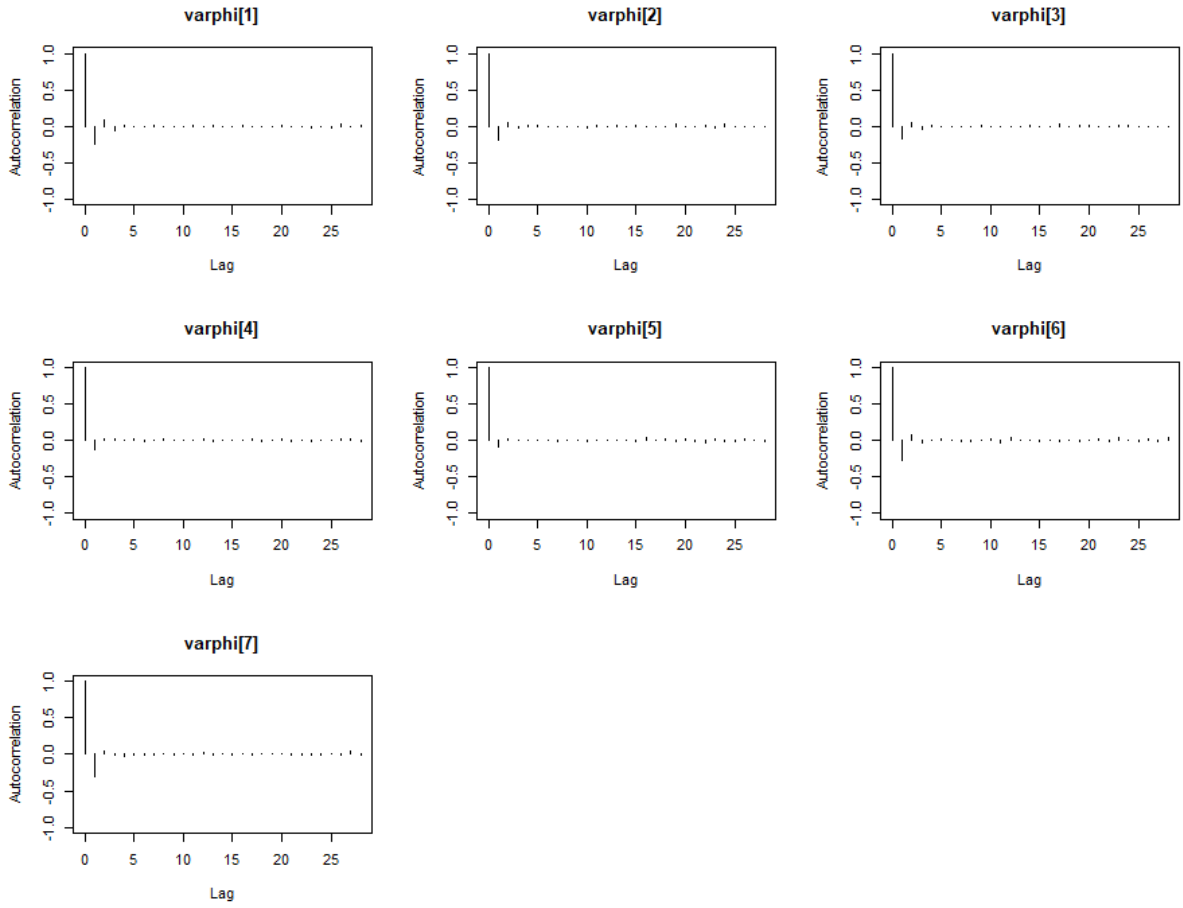
Já as Figuras 6.7, 6.8 e 6.9 apresentam os gráficos de trajetória e das curvas de densidade, autocorrelação e Gelman-Rubin, respectivamente, relacionados ao vetor de parâmetros $\boldsymbol{\varphi}$, associado aos preditores observados do desfecho distal. O número de parâmetros no vetor $\boldsymbol{\varphi}$ está de acordo com a parametrização usando casela de referência para a inclusão dos preditores observados no modelo de riscos proporcionais de Cox. Desta forma, o parâmetro φ_1 representa o logaritmo da razão das funções taxa de falha para indivíduos com idade de 18 a 19 anos se comparados com indivíduos com idade de 15 a 17 anos. Interpretação análoga é feita com os demais parâmetros e níveis dos preditores categóricos observados conforme apresentado na Tabela 6.4.

Figura 6.7: *Traceplots* e curvas de densidade para os parâmetros do vetor φ, de acordo com 5 cadeias geradas pelo método BS para os dados PrEP. 2019-2021



A Figura 6.7 apresenta o *traceplot* e a curva de densidade associados aos parâmetros individuais do vetor de parâmetros $\boldsymbol{\varphi}$. Para todos os parâmetros percebe-se que as cadeias apresentadas misturam-se adequadamente, não sendo identificadas tendências entre as cadeias exploradas, com ambas cadeias para cada um dos parâmetros apresentando valores em torno de uma média estável. Além disto, todos os gráficos de densidade da distribuição a *posteriori* para os parâmetros do vetor $\boldsymbol{\varphi}$ apresentam apenas uma única moda, estão distribuídos simetricamente em torno de um valor e tem caudas curtas, refletindo numa maior assertividade no cálculo das estimativas dos parâmetros associados ao modelo semiparamétrico de Cox.

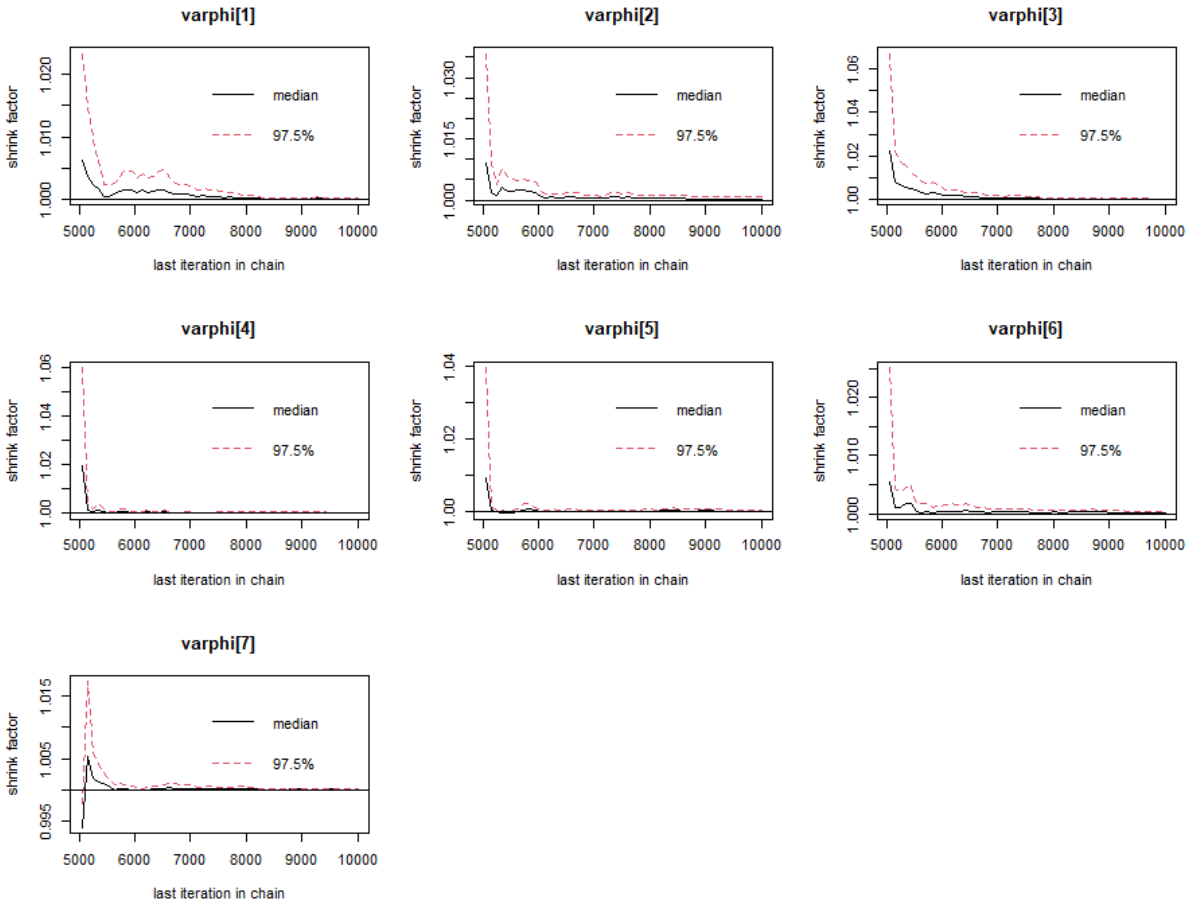
Figura 6.8: Autocorrelação para valores amostrados dos parâmetros $\boldsymbol{\varphi}$, de acordo com uma cadeia, gerada pelo método BS para os dados PrEP. 2019-2021



A Figura 6.8 apresenta os gráficos de autocorrelação nos valores amostrados para os parâmetros do vetor $\boldsymbol{\varphi}$. Para cada um dos parâmetros nota-se que a partir do *lag* 3 não são detectadas correlações expressivas entre as amostras a posteriori para $\boldsymbol{\varphi}$, o que é um indicativo de que a autocorrelação diminui rapidamente com o aumento do *lag* para ambos os parâmetros associados às variáveis preditoras, sendo um indicativo de que as amostras a posteriori são independentes. O mesmo tipo de gráfico para as demais cadeias

apresenta comportamento análogo ao gráfico com dados da primeira cadeia.

Figura 6.9: Gráfico de estatística Gelman-Rubin para os parâmetros do vetor $\boldsymbol{\varphi}$, de acordo com 5 cadeias geradas pelo método BS para os dados PrEP. 2019-2021



A Figura 6.9 apresenta o diagnóstico de Gelman-Rubin para os parâmetros do vetor $\boldsymbol{\varphi}$, avaliando se as diferentes cadeias convergiram para a mesma distribuição, considerando cada um dos parâmetros. A partir do gráfico tem-se evidências de que as cadeias convergem para a mesma distribuição *a posteriori* após 5000 iterações para os parâmetros φ_4 e φ_5. Para os demais parâmetros, foi observado número de iterações superior a 6000, uma vez que o fator de redução de escala em potencial se aproxima do valor 1.

As Figuras 6.7, 6.8 e 6.9 fornecem evidências de convergência das cadeias, independência das distribuições *a posteriori* e número observado de iterações até que se atinja a convergência. Desta forma, há evidências de que o processo de estimação e convergência dos parâmetros do submodelo estrutural ocorreu de forma adequada. Contudo, para que seja corroborada a estimativa do parâmetro β, é necessário também realizar

procedimento de diagnóstico de convergência para os parâmetros do submodelo estrutural de LCA, de acordo com o método de estimação utilizado na análise (método BS). No Apêndice A são apresentados os gráficos adicionais para avaliação de convergência, referentes aos parâmetros do submodelo de mensuração e parâmetros associados à taxa de falha basal do submodelo estrutural, segundo método de estimação BS, para análise dos dados do projeto PrEP1519. Através da análise destes gráficos não são detectados problemas relacionados à convergência dos demais parâmetros do modelo, o que garante a validade das estimativas obtidas para os dados estudados.

Capítulo 7

Considerações Finais

Embora a formalização dos modelos para estimação do efeito de variáveis latentes em uma resposta distal, sob a ótica frequentista, esteja disponível na literatura, os correspondentes métodos computacionais estatísticos não estão implementados ou acessíveis para os demais pesquisadores que desejem utilizar esta metodologia. Atualmente, os métodos disponíveis para estimação do efeito de uma variável latente em um desfecho distal, ao utilizar abordagens do tipo classificação e análise, consideram estratégias para atenuar o viés das estimativas relacionadas aos preditores latentes, como apresentado na Seção 2.3. Como discutido em Bakk, Tekle e Vermunt (2013), o uso da abordagem de três etapas não corrigida leva a estimativas de parâmetros seriamente tendenciosas referentes à associação de pertencimento à classe latente com variáveis externas. Entretanto, mesmo a incorporação de métodos de correção pode ocasionar em subestimação das estimativas dos parâmetros e respectivos erros-padrão, particularmente em cenários em que a separação de classes do modelo de mensuração é muito baixa. Em Lanza, Tan e Bray (2013) também é proposta uma abordagem para eliminação do viés causado pelo uso de abordagens do tipo classificação e análise na estimação de parâmetros associados a preditores latentes de um desfecho distal. Contudo, esta metodologia apresenta limitações uma vez que considera inicialmente a inclusão do desfecho latente como uma covariável do modelo de mensuração, o que torna inviável sua aplicação em análises com desfechos censurados. Adicionalmente, as metodologias mais disponíveis na literatura não consideram a inclusão de outras variáveis observadas como preditoras do desfecho observado.

Neste trabalho os métodos BSM e BS, propostos por Costa, Amorim e Bispo (2021), foram estendidos e adaptados ao contexto em que um desfecho distal é definido pelo tempo até a ocorrência de um evento. O método BSM, baseado na abordagem de classificação e análise, foi implementado para que fosse possível realizar uma comparação com o método BS que realiza a estimação simultânea dos parâmetros do modelo. Estes métodos foram desenvolvidos sob a perspectiva bayesiana e, desta forma, permitem que informações *a*

priori sejam incorporadas ao modelo de forma a diminuir a incerteza dos parâmetros estimados, além de facilitar a obtenção de estimativas intervalares para os parâmetros do submodelo de mensuração. Neste trabalho foram consideradas distribuições *a priori* não informativas para os parâmetros estimados, em ambos os métodos. Além disto, para ambos os métodos é possível realizar a incorporação de variáveis observadas como preditoras do desfecho distal, além da variável latente.

Os resultados dos estudos de simulação realizados no Capítulo 5 evidenciaram que o Método Bayesiano Modal Simplificado (BSM), ao ser utilizado como abordagem para estimação dos parâmetros associados ao modelo com desfecho distal, resultou em estimativas enviesadas para o parâmetro associado à variável latente em todos os cenários abordados. Ao analisar o viés relativo percentual associado à estimativa do parâmetro β_c, observa-se uma subestimação ao utilizar o método BSM. Nos cenários com alta entropia, as estimativas são 16 a 18% menores, enquanto nos cenários com baixa entropia, essa subestimação varia entre 33 e 40% em relação aos valores verdadeiros do parâmetro. Em relação à variância e raiz do erro quadrático médio das estimativas, estas reduziram com o aumento do tamanho amostral, contudo, a probabilidade de cobertura percentual apresentou reduções atenuadas com o aumento do tamanho amostral nos cenários com entropia baixa. Como apresentado no Capítulo 5 o aumento do tamanho amostral ocasiona a diminuição da variabilidade das estimativas, resultando em intervalos de credibilidade mais concentrados em torno de estimativas enviesadas, que chegam a ser 40% menores nos cenários com entropia baixa no método BSM. Desta forma há uma redução no $CP_{95\%}(\beta_c)$ mais acentuada nestes cenários à medida que o tamanho amostral aumenta.

O método Método Bayesiano Simultâneo (BS), por sua vez, apresentou resultados satisfatórios em todos os cenários analisados, embora tenha mostrado estimativas ligeiramente mais enviesadas nos casos de menor tamanho amostral, baixa entropia e maior percentual de censura. Além disso, observou-se uma diminuição na variância e na raiz do erro quadrático médio das estimativas conforme o tamanho amostral aumentava. Diferente do método anterior, a probabilidade de cobertura percentual não apresentou reduções significativas com o aumento do tamanho amostral nos cenários de baixa entropia. O método BS também teve bom desempenho ao lidar com uma maior complexidade do modelo, incluindo preditores observados para o desfecho distal. Mesmo em cenários com baixa entropia e maior percentual de censura, o método BS apresentou baixo viés com tamanhos amostrais maiores em comparação com o método BSM.

Na análise dos dados do projeto PrEP1519, os métodos BS e BSM foram ajustados para permitir a comparação, especialmente em relação à estimativa do parâmetro associado à variável latente que representa o perfil de risco real ao HIV dos participantes. Utilizando critérios de seleção de modelos no contexto da análise de classes latentes, foi

selecionado um submodelo de mensuração LCA com 2 classes. Ambos os métodos produziram estimativas próximas para os parâmetros do submodelo de mensuração, como esperado, pois foram utilizadas prioris vagas para os parâmetros e, teoricamente, o submodelo de mensuração não deveria ser influenciado pelo submodelo estrutural. No submodelo estrutural, tanto para o método BS quanto para o método BSM, o risco de primeira descontinuidade do tratamento PrEP foi menor para o grupo com perfil de alto risco em comparação com o grupo de baixo risco. No entanto, conforme indicado pela literatura e confirmado pelos estudos de simulação deste trabalho, o método BSM parece subestimar o efeito associado ao preditor latente. Enquanto o método BS indicou uma redução de 41% no risco de descontinuidade para indivíduos de alto risco, o método BSM apresentou uma redução de apenas 19%.

Uma importante consideração a ser destacada na implementação dos métodos BS e BSM, particularmente em amostras com número elevado de indivíduos, maior percentual de censura e entropia baixa, é o tempo gasto para a convergência do processo de estimação dos parâmetros. Considerando a estimação sequencial dos modelos para as amostras provenientes do estudo de simulação, o tempo para estimação de 1.000 modelos aumenta substancialmente, sendo necessário aproximadamente uma semana para estimação em cenários mais simples utilizando a metodologia proposta via STAN. Desta forma, o uso de técnicas de computação paralela se faz indispensável para a realização de estudos de simulações em tempo viável.

O método BS, apesar de demandar mais tempo computacional em comparação ao BSM, mostrou-se uma alternativa viável e não tendenciosa, sem a necessidade de grandes amostras para obter resultados consistentes, especialmente em cenários com maior entropia. Além disso, é possível investigar a inclusão de prioris informativas na abordagem bayesiana para avaliar o viés em todos os métodos considerados. Para trabalhos futuros, sugere-se a extensão do estudo para outros tipos de submodelos estruturais em desfechos de tempo até o evento, como os modelos paramétricos baseados nas distribuições Exponencial, Weibull e Gompertz, além dos modelos de tempo de falha acelerado (AFT). Outra recomendação é o uso de *B-splines* para incorporar preditores observados dependentes do tempo no submodelo estrutural.

Bibliografia

Akaike, H. (1998). Information Theory and an Extension of the Maximum Likelihood Principle. Em E. Parzen, K. Tanabe & G. Kitagawa (Eds.), *Selected Papers of Hirotugu Akaike* (pp. 199–213). Springer New York. https://doi.org/10.1007/978-1-4612-1694-0_15

Amorim, L. D. A. F., Fiaccone, R. L., Santos, C. A. S., Santos, T. N. d., de Moraes, L. T. L., Oliveira, N. F., Barbosa, S. O., Santos, D. N. d., Santos, L. M. d., & Matos, S. (2010). Structural equation modeling in epidemiology. *Cadernos de Saúde Pública*, *26*, 2251–2262.

Andersen, P., Borgan, 0., Gill, R., & Keiding, N. (1993). Statistical Models Based on Counting Processes. *Springer New York, Ser. Statist.*

Asparouhov, T., Masyn, K., & Muthen, B. (2006). Continuous time survival in latent variable models. *Proceedings of the Joint Statistical Meeting in Seattle*, 180–187.

Asparouhov, T., & Muthén, B. (2011). Using Bayesian priors for more flexible latent class analysis. *Proceedings of the 2011 Joint Statistical Meeting, Miami Beach, FL.*

Asparouhov, T., & Muthén, B. (2014a). Auxiliary variables in mixture modeling: Three-step approaches using M plus. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(3), 329–341.

Asparouhov, T., & Muthén, B. (2014b). Auxiliary variables in mixture modeling: Using the BCH method in Mplus to estimate a distal outcome model and an arbitrary secondary model. *Mplus Web Notes*, *21*(2), 1–22.

Azuero, A. (2016). A note on the magnitude of hazard ratios. *Cancer*, *122*(8), 1298–1299.

Bakk, Z., & Kuha, J. (2017). Two-step estimation of models between latent classes and external variables. *Psychometrika*, *83*, 871–892.

Bakk, Z., & Kuha, J. (2018). Two-step estimation of models between latent classes and external variables. *Psychometrika*, *83*, 871–892.

Bakk, Z., & Kuha, J. (2021). Relating latent class membership to external variables: An overview. *British Journal of Mathematical and Statistical Psychology*, *74*(2), 340–362.

Bakk, Z., Oberski, D. L., & Vermunt, J. K. (2016). Relating latent class membership to continuous distal outcomes: Improving the LTB approach and a modified three-step implementation. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(2), 278–289.

Bakk, Z., Tekle, F. B., & Vermunt, J. K. (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological Methodology*, *43*(1), 272–311.

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., & Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, *92*(440), 1375–1386.

Barbosa, H., De Oliveira, F., Dos Santos, T., Fonseca, G., & Monteiro, J. (2010). Cálculo de Média a posteriori através de Métodos de Integração Numérica e Simulação Monte Carlo: Estudo Comparativo. *Revista INGEPRO-Inovação, Gestão e Produção*, *2*, 60–74.

Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics In Medicine*, *24*(11), 1713–1723.

Bengtsson, H. (2021). A Unifying Framework for Parallel and Distributed Processing in R using Futures. *The R Journal*, *13*(2), 208–227. https://doi.org/10.32614/RJ-2021-048

Betancourt, M., & Girolami, M. (2015). Hamiltonian Monte Carlo for hierarchical models. *Current Trends in Bayesian Methodology with Applications*, *79*(30), 2–4.

Bispo, G. S. (2019). *Inferência em Modelos com Respostas Distais: Uma Abordagem Bayesiana* [Master's thesis]. Universidade Federal da Bahia - Instituto de Matemática e Estatística.

Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*, *12*(1), 3–27.

Bray, B. C., Lanza, S. T., & Tan, X. (2015). Eliminating bias in classify-analyze approaches for latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(1), 1–11.

Brilleman, S. L., Elci, E. M., Novik, J. B., & Wolfe, R. (2020). Bayesian survival analysis using the rstanarm R package. *arXiv preprint arXiv:2002.09633*. https://arxiv.org/pdf/2002.09633

Brooks, S., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455.

Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. Chapman; Hall/CRC. https://doi.org/10.1201/b10905

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

Bürkner, P.-C. (2021). Bayesian Item Response Modeling in R with brms and Stan. *Journal of Statistical Software*, *100*(5), 1–54. https://doi.org/10.18637/jss.v100.i05

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, *13*, 195–212.

Chen, Q., Wu, H., Ware, L. B., & Koyama, T. (2014). A Bayesian approach for the Cox proportional hazards model with covariates subject to detection limit. *International Journal of Statistics in Medical Research*, *3*(1), 32.

Clark, S. L., & Muthén, B. (2009). Latent Class Analysis Results to Variables not Included in the Analysis. https://api.semanticscholar.org/CorpusID:6401442

Collett, D. (1994). *Modelling survival data in medical research*. CRC press.

Collins, L. M., & Lanza, S. T. (2009). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Vol. 718). John Wiley & Sons.

Colosimo, E. A., & Giolo, S. R. (2006). *Análise de sobrevivência aplicada*. Editora Blucher.

Costa, L. C. d., Amorim, L. D., & Bispo, G. S. (2021). Bayesian Computation via the Gibbs Sampler for Mixture Models with Gaussian Distal Outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(6), 839–850.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202.

Curry, H. B., & Schoenberg, I. J. (1966). On Pólya frequency functions IV: the fundamental spline functions and their limits. *J. Analyse Math*, *17*(71), 107.

Darney, D., Reinke, W. M., Herman, K. C., Stormont, M., & Ialongo, N. S. (2013). Children with co-occurring academic and behavior problems in first grade: Distal outcomes in twelfth grade. *Journal of School Psychology*, *51*(1), 117–128.

Dayton, C. M., & Macready, G. B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, *83*(401), 173–178.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Dourado, I., Magno, L., Greco, D. B., Zucchi, E. M., Ferraz, D., Westin, M. R., & Grangeiro, A. (2023). Interdisciplinarity in HIV prevention research: the experience of the PrEP1519 study protocol among adolescent MSM and TGW in Brazil. *Cadernos de Saúde Pública*, *39*, e00143221.

Dourado, I., Soares, F., Magno, L., Amorim, L., Eustorgio Filho, M., Leite, B., Greco, D., Westin, M., Tupinambás, U., Massa, P., Zucchi, E. M., & Grangeiro, A. (2023). Adherence, safety, and feasibility of HIV pre-exposure prophylaxis among adolescent men who have sex with men and transgender women in Brazil (PrEP1519 Study). *Journal of Adolescent Health*, *73*(6), S33–S42.

Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid monte carlo. *Physics letters B*, *195*(2), 216–222.

Friedman, M. (1982). Piecewise exponential models for survival data with covariates. *The Annals of Statistics*, *10*(1), 101–113.

Gelfand, A. E., & Mallick, B. K. (1995). Bayesian analysis of proportional hazards models built from monotone functions. *Biometrics*, 843–852.

Gelman, A., Lee, D., & Guo, J. (2015). Stan: A probabilistic programming language for Bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, *40*(5), 530–543.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6), 721–741.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. (1996). *Markov Chain Monte Carlo in Practice*. CRC Press.

Grant, R. M., Lama, J. R., Anderson, P. L., McMahan, V., Liu, A. Y., Vargas, L., Goicochea, P., Casapía, M., Guanira-Carranza, J. V., Ramirez-Cardich, M. E., et al. (2010). Preexposure chemoprophylaxis for HIV prevention in men who have sex with men. *New England Journal of Medicine*, *363*(27), 2587–2599.

Hagenaars, J. A. (1990). *Categorical Longitudinal Data: Log-linear panel, trend, and cohort analysis*. SAGE Publications.

Heinzl, H., & Kaider, A. (1997). Gaining more flexibility in Cox proportional hazards regression models with cubic spline functions. *Computer Methods and Programs in Biomedicine*, *54*(3), 201–208.

Ibrahim, J. G., Chen, M.-H., & Sinha, D. (2001). *Bayesian survival analysis* (Vol. 2). Springer.

Işık, H., Karasoy, D., & Karabey, U. (2023). A new adjusted Bayesian method in Cox regression model with covariate subject to measurement error. *Hacettepe Journal of Mathematics and Statistics*, *52*(5), 1367–1378.

Klein, J. P., & Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data* (Vol. 1230). Springer.

Klein, J. P., Van Houwelingen, H. C., Ibrahim, J. G., & Scheike, T. H. (2013). *Handbook of Survival Analysis* (1st). CRC Press. https://doi.org/https://doi.org/10.1201/b16248

Lanza, S. T., Collins, L. M., Lemmon, D. R., & Schafer, J. L. (2007). PROC LCA: A SAS procedure for Latent Class Analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 671–694.

Lanza, S. T., Tan, X., & Bray, B. C. (2013). Latent Class Analysis with distal outcomes: A flexible model-based approach. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(1), 1–26.

Larsen, K. (2004). Joint analysis of time-to-event and multiple binary indicators of latent classes. *Biometrics*, *60*(1), 85–92.

Lee, S.-Y. (2007). *Structural Equation Modeling: A Bayesian Approach*. John Wiley & Sons.

Leonard, T. (1978). Density estimation, stochastic processes and prior information. *Journal of the Royal Statistical Society: Series B (Methodological)*, *40*(2), 113–132.

Li, Y., Lord-Bessen, J., Shiyko, M., & Loeb, R. (2018). Bayesian Latent Class Analysis tutorial. *Multivariate Behavioral Research*, *53*(3), 430–451.

Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R package for polytomous variable Latent Class Analysis. *Journal of Statistical Software*, *42*, 1–29.

Lythgoe, D. T., Garcia-Fiñana, M., & Cox, T. F. (2019). Latent class modeling with a time-to-event distal outcome: A comparison of one, two and three-step approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(1), 51–65.

McCormack, S., Dunn, D. T., Desai, M., Dolling, D. I., Gafos, M., Gilson, R., Sullivan, A. K., Clarke, A., Reeves, I., Schembri, G., et al. (2016). Pre-exposure prophylaxis to prevent the acquisition of HIV-1 infection (PROUD): effectiveness results from the pilot phase of a pragmatic open-label randomised trial. *The Lancet*, *387*(10013), 53–60.

Merkle, E. C., Fitzsimmons, E., Uanhoro, J., & Goodrich, B. (2020). Efficient Bayesian Structural Equation Modeling in Stan. *arXiv preprint arXiv:2008.07733*.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092.

Monnahan, C. C., Thorson, J. T., & Branch, T. A. (2017). Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution*, *8*(3), 339–348.

Morin, R. T., Insel, P., Nelson, C., Butters, M., Bickford, D., Landau, S., Saykin, A., Weiner, M., & Mackin, R. S. (2019). Latent classes of cognitive functioning among

depressed older adults without dementia. *Journal of the International Neuropsychological Society*, *25*(8), 811–820.

Muthén, B. (2004). Latent variable analysis. *The Sage Handbook of Quantitative Methodology for the Social Sciences*, *345*(368), 106–109.

Muthén, B., & Asparouhov, T. (2009). Multilevel regression mixture analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(3), 639–657.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus Version 8 User's Guide*. Los Angeles, CA: Muthén & Muthén. https://www.statmodel.com/download/usersguide/MplusUserGuideVer_8.pdf

Nagin, D. (2005). Group-based modeling of development Harvard University Press. *Cambridge, Mass*, *4159*, 9780674041318.

Neal, R. M. (2011). Chapter 5: MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, *2*(11), 2.

Nylund, K., Bellmore, A., Nishina, A., & Graham, S. (2007). Subtypes, severity, and structural stability of peer victimization: What does latent class analysis say? *Child Development*, *78*(6), 1706–1722.

O'Hagan, A. (1994). *Bayesian Statistics. Kendall's Advanced Theory of Statistics* (Vol. 2B). Cambridge University Press, Cambridge.

Olino, T. M., Klein, D. N., Lewinsohn, P. M., Rohde, P., & Seeley, J. R. (2010). Latent trajectory classes of depressive and anxiety disorders from adolescence to adulthood: descriptions of classes and associations with risk factors. *Comprehensive Psychiatry*, *51*(3), 224–235.

Ormerod, J. T., & Wand, M. P. (2010). Explaining variational approximations. *The American Statistician*, *64*(2), 140–153.

Palomo, J., Dunson, D. B., & Bollen, K. (2007). 8 - Bayesian Structural Equation Modeling. Em S.-Y. Lee (Ed.), *Handbook of Latent Variable and Related Models* (pp. 163–188). North-Holland. https://doi.org/https://doi.org/10.1016/B978-044452044-9/50011-2

Plummer, M. (2017). JAGS: Just another Gibbs sampler. http://mcmc-jags.sourceforge.net/

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News*, *6*(1), 7–11. https://journal.r-project.org/archive/

Raftery, A. E., & Lewis, S. (1992). How many iterations in the Gibbs sampler. *Bayesian Statistics*, *4*(2), 763–773.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science*, 425–441.

Redner, R. A., & Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, *26*(2), 195–239.

Roberts, T. J., & Ward, S. E. (2011). Using Latent Transition Analysis in nursing research to explore change over time. *Nursing Research*, *60*(1), 73.

Rosenberg, P. S. (1995). Hazard function estimation using B-splines. *Biometrics*, 874–887.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys* (Vol. 81). John Wiley & Sons.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 461–464.

Sinha, D. (1993). Semiparametric Bayesian analysis of multiple event time data. *Journal of the American Statistical Association*, *88*(423), 979–983.

Skinner, C. J., Holt, D., & Smith, T. F. (1989). *Analysis of complex surveys*. John Wiley & Sons Ltd.

Sleeper, L. A., & Harrington, D. P. (1990). Regression splines in the Cox model with application to covariate effects in liver disease. *Journal of the American Statistical Association*, *85*(412), 941–949.

Smid, S. C., Depaoli, S., & Van De Schoot, R. (2020). Predicting a distal outcome variable from a latent growth model: ML versus Bayesian estimation. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(2), 169–191.

Smith, A. F., & Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, *55*(1), 3–23.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639.

Stan Development Team. (2024a). RStan: the R interface to Stan [R package version 2.32.6]. https://mc-stan.org/

Stan Development Team. (2024b). Stan Reference Manual, Version 2.35. https://mc-stan.org/docs/reference-manual/

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(4), 795–809.

Therneau, P., & Grambsch, T. (2000). *Modeling survival data: extending the Cox model*. Springer, New York, NY.

van Lang, N. D., Ferdinand, R. F., Ormel, J., & Verhulst, F. C. (2006). Latent Class Analysis of anxiety and depressive symptoms of the Youth Self-Report in a general population sample of young adolescents. *Behaviour Research and Therapy*, *44*(6), 849–860.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, *18*(4), 450–469.

Wang, C., Brown, H. C., & Bandeen-Roche, K. (2005). Residual diagnostics for growth mixture models: Examining the impact of a preventive intervention on multiple trajectories of aggressive behavior. *Journal of the American Statistical Association*, *100*(471), 1054–1076.

Wang, W., & Yan, J. (2021). Shape-Restricted Regression Splines with R Package splines2. *Journal of Data Science*, *19*(3).

Wang, W., & Yan, J. (2024). *splines2: Regression Spline Functions and Classes* [R package version 0.5.2]. https://CRAN.R-project.org/package=splines2

White, A., & Murphy, T. B. (2014). BayesLCA: An R package for Bayesian latent class analysis. *Journal of Statiscal Software*, *61*(13), 1–28.

White, I., & Royston, P. (2009). Imputing missing covariate values for the Cox model. *Statistics in Medicine*, *28*(15), 1982–1998.

Xue, Q.-L., & Bandeen-Roche, K. (2004). Combining complete multivariate outcomes with incomplete covariate information: a latent class approach. *Biometrics*, *58*(1), 110–120. https://doi.org/10.1111/j.0006-341X.2002.00110.x

Yamaguchi, K. (2000). Multinomial logit latent-class regression models: An analysis of the predictors of gender-role attitudes among Japanese women. *American Journal of Sociology*, *105*(6), 1702–1740.

Zeballos, D., Magno, L., Soares, F., Eustorgio Filho, M., Amorim, L., Pinto Jr, J. A., Greco, D., Grangeiro, A., & Dourado, I. (2023). Oral pre-exposure prophylaxis for HIV discontinuation in a large cohort of adolescent men who have sex with men and transgender women in Brazil. *Journal of Adolescent Health*, *73*(6), S43–S49.

Zhang, J., Li, C., Xu, J., Hu, Z., Rutstein, S. E., Tucker, J. D., Ong, J. J., Jiang, Y., Geng, W., Wright, S. T., et al. (2022). Discontinuation, suboptimal adherence, and reinitiation of oral HIV pre-exposure prophylaxis: a global systematic review and meta-analysis. *The Lancet HIV*, *9*(4), e254–e268.

Apêndice A

Resultados complementares

A.1 Gráficos suplementares para avaliação de convergência na análise dos dados PrEP1519

Figura A.1: *Traceplot* e curva de densidade para os parâmetros $\boldsymbol{\gamma}$, de acordo com 5 cadeias geradas pelo método BS para os dados PrEP. 2019-2021

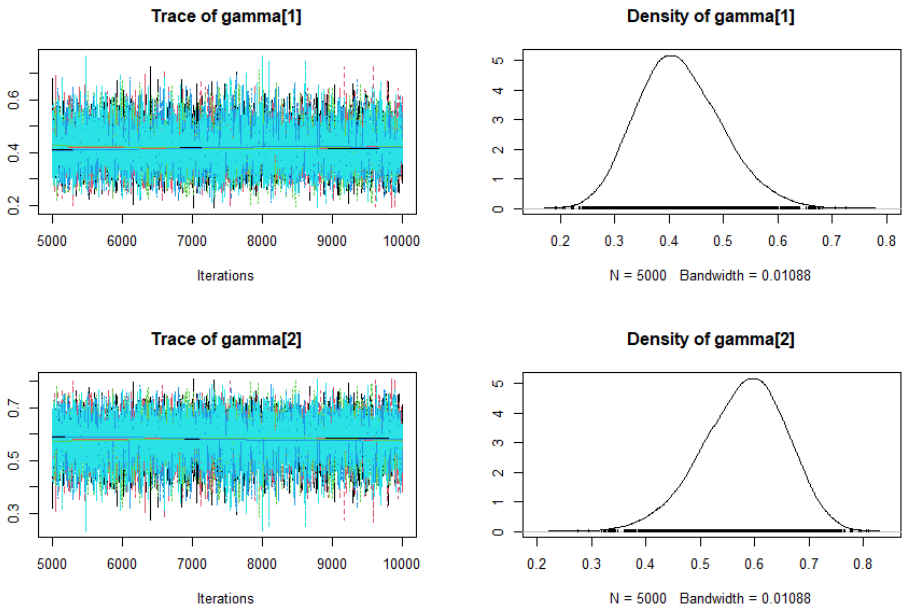


Figura A.2: Autocorrelação para valores amostrados dos parâmetros $\boldsymbol{\gamma}$, de acordo com uma cadeia, gerada pelo método BS para os dados PrEP. 2019-2021

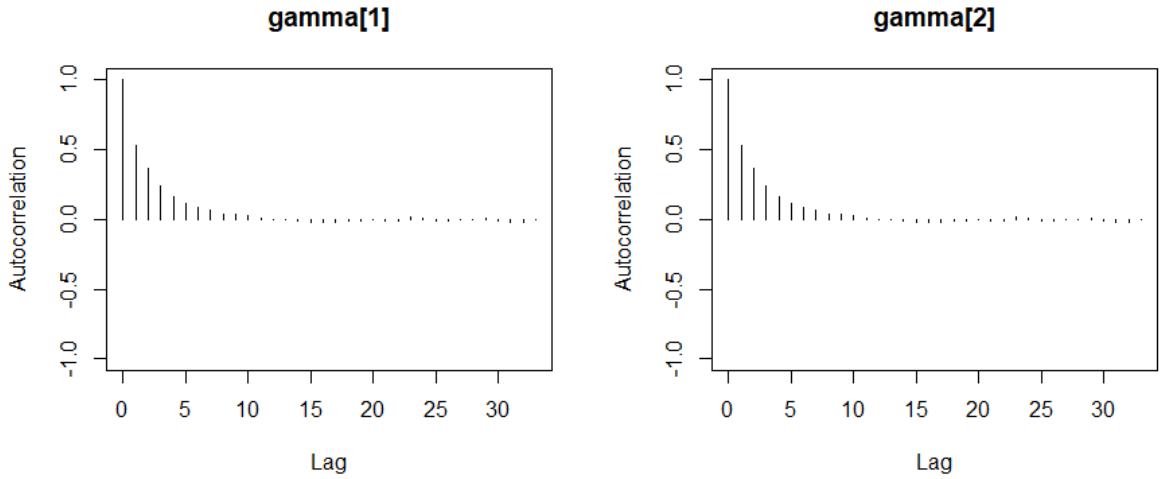


Figura A.3: Gráfico de estatística Gelman-Rubin para os parâmetros $\boldsymbol{\gamma}$, de acordo com 5 cadeias geradas pelo método BS para os dados PrEP. 2019-2021

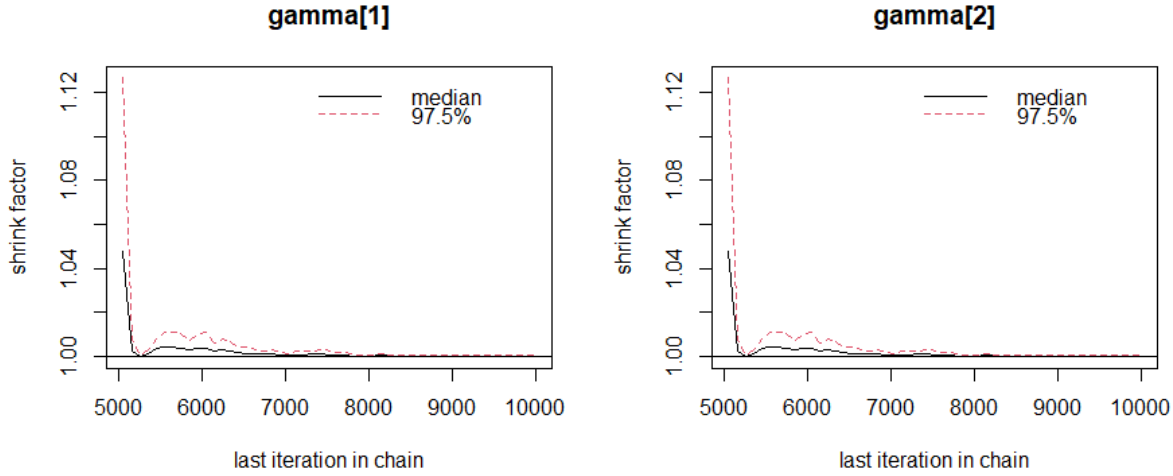


Figura A.4: *Traceplots* e curvas de densidade para os parâmetros $\boldsymbol{\rho_{k,r_k|c=1}}$, de acordo com 5 cadeias geradas pelo método BS para os dados PrEP. 2019-2021

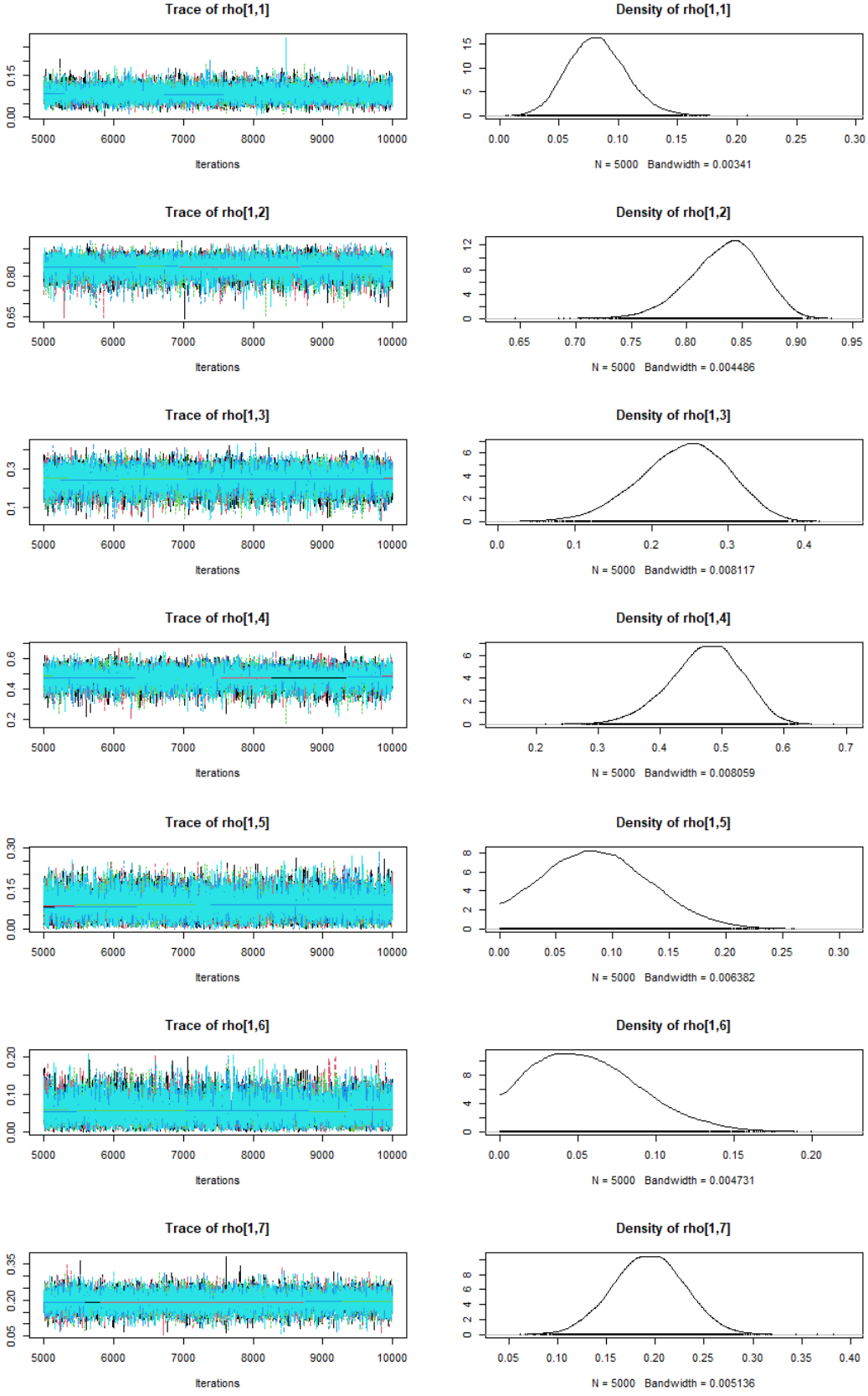


Figura A.5: Autocorrelação para valores amostrados dos parâmetros $\boldsymbol{\rho_{k,r_k|c=1}}$, de acordo com uma cadeia, gerada pelo método BS para os dados PrEP. 2019-2021

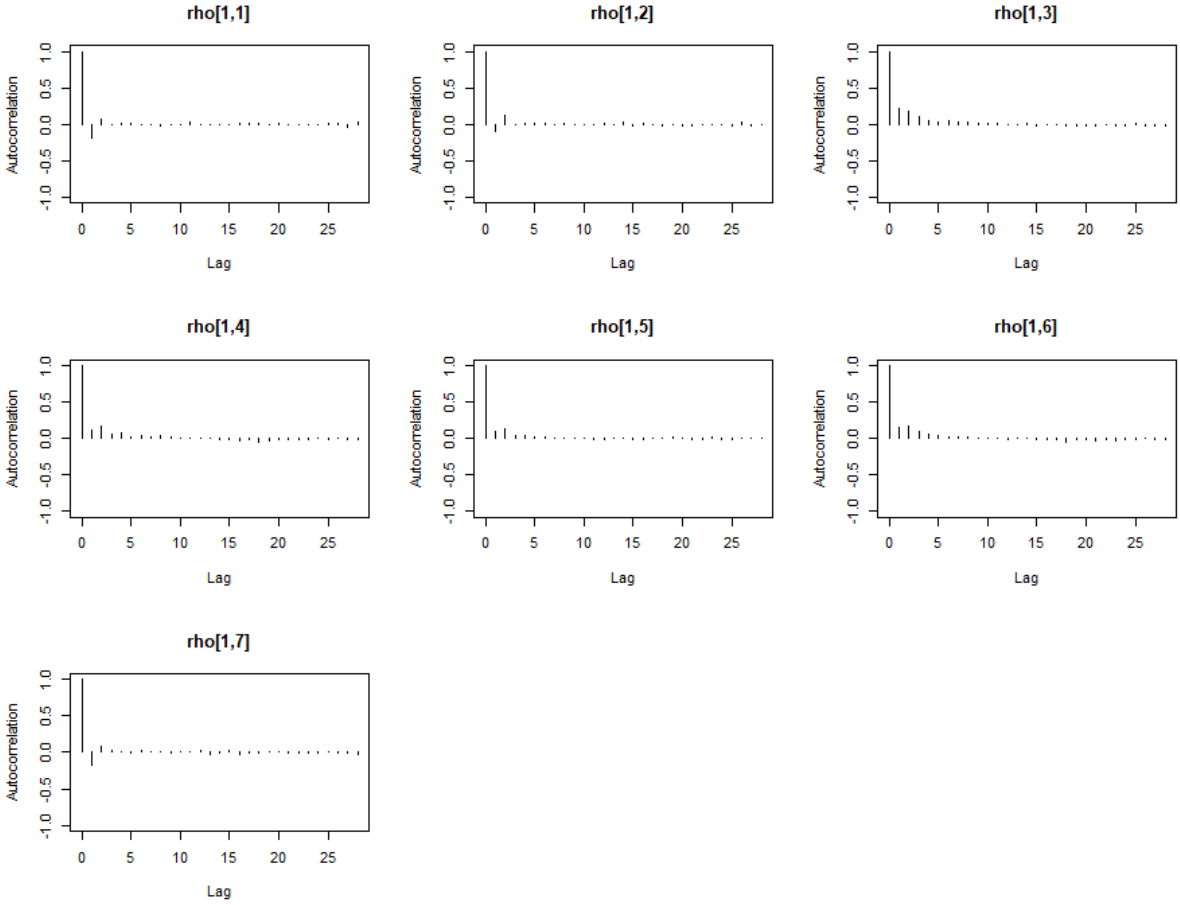


Figura A.6: Gráfico de estatística Gelman-Rubin para os parâmetros $\boldsymbol{\rho_{k,r_k|c=1}}$, de acordo com 5 cadeias geradas pelo método BS para os dados PrEP. 2019-2021

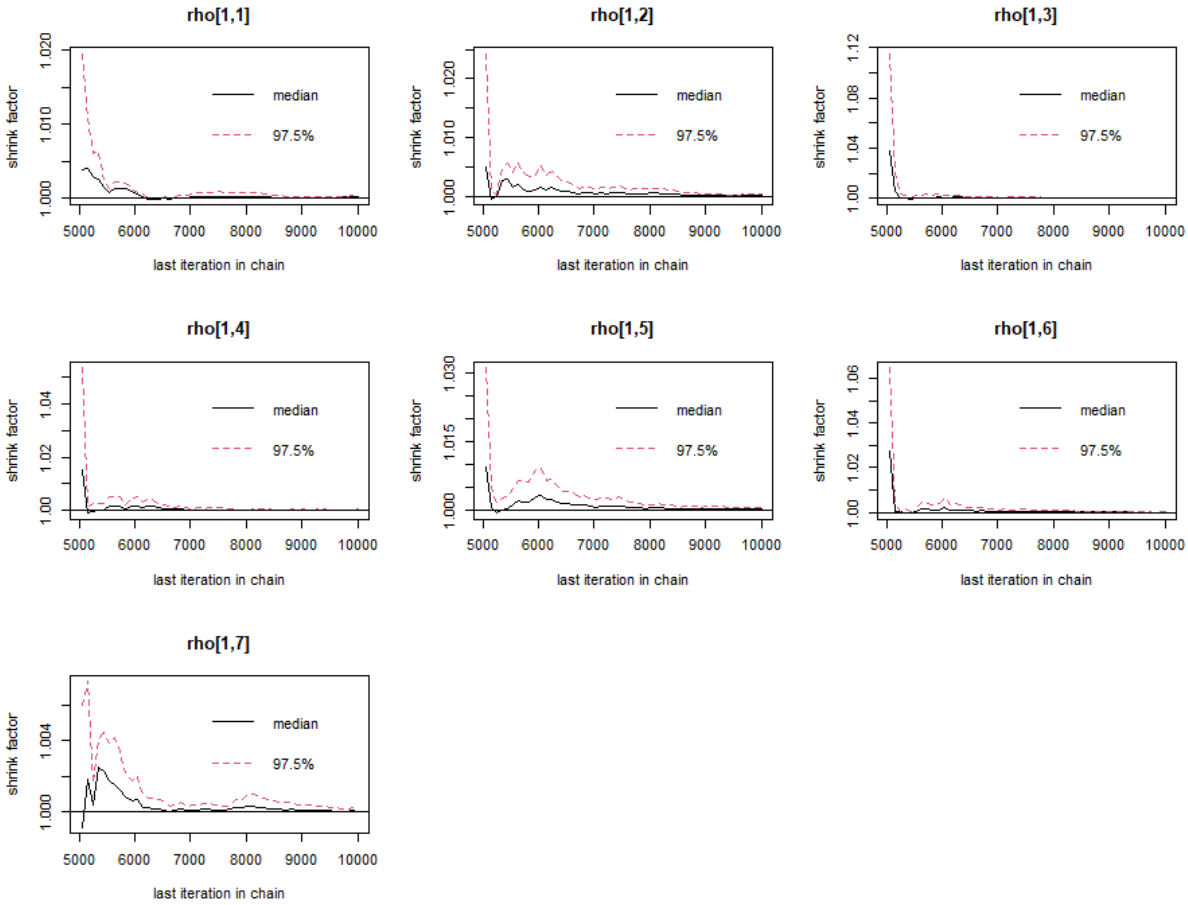


Figura A.7: *Traceplots* e curvas de densidade para os parâmetros $\boldsymbol{\rho_{k,r_k|c=2}}$, de acordo com 5 cadeias geradas pelo método BS para os dados PrEP. 2019-2021

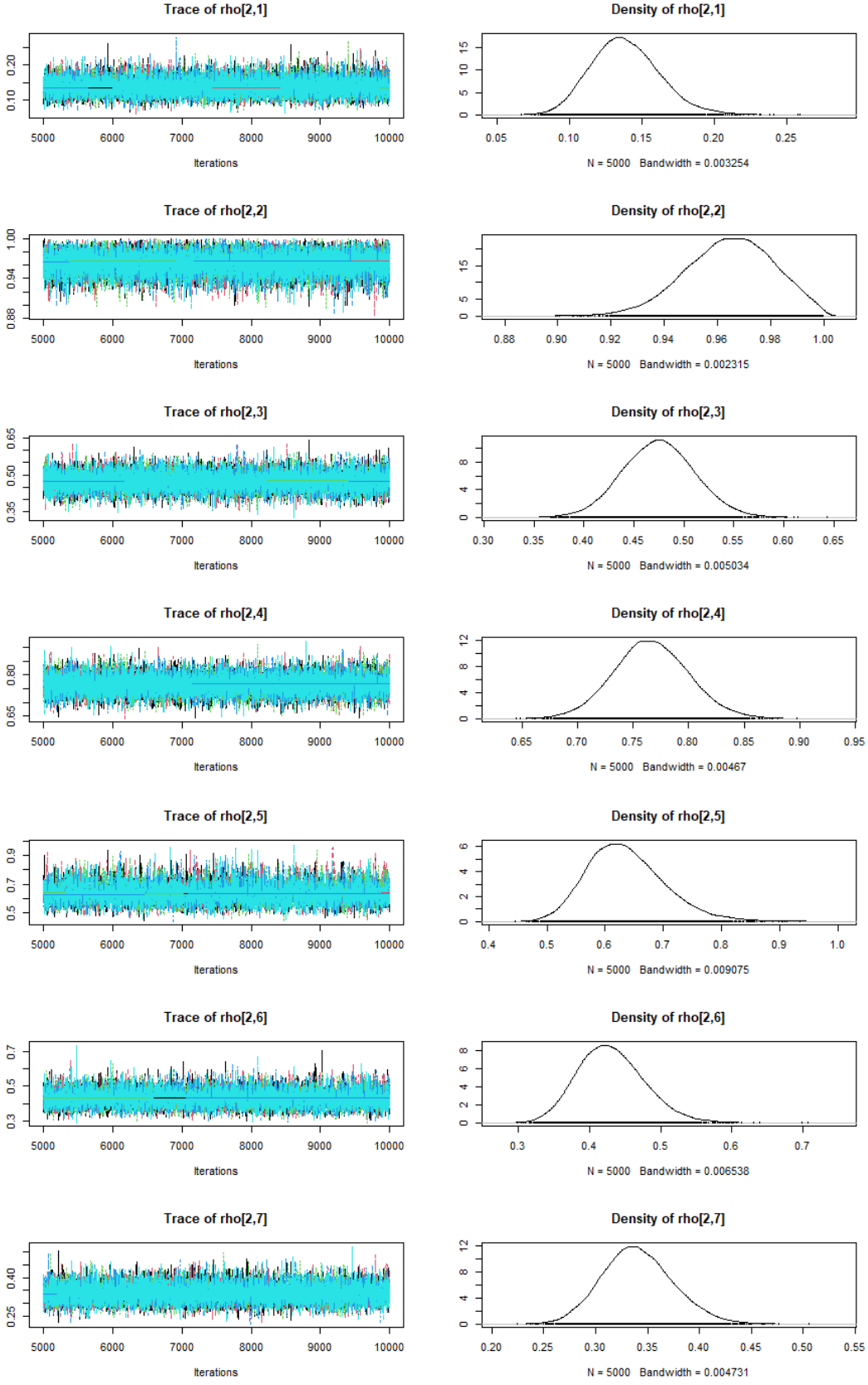


Figura A.8: Autocorrelação para valores amostrados dos parâmetros $\boldsymbol{\rho_{k,r_k|c=2}}$, de acordo com uma cadeia, gerada pelo método BS para os dados PrEP. 2019-2021

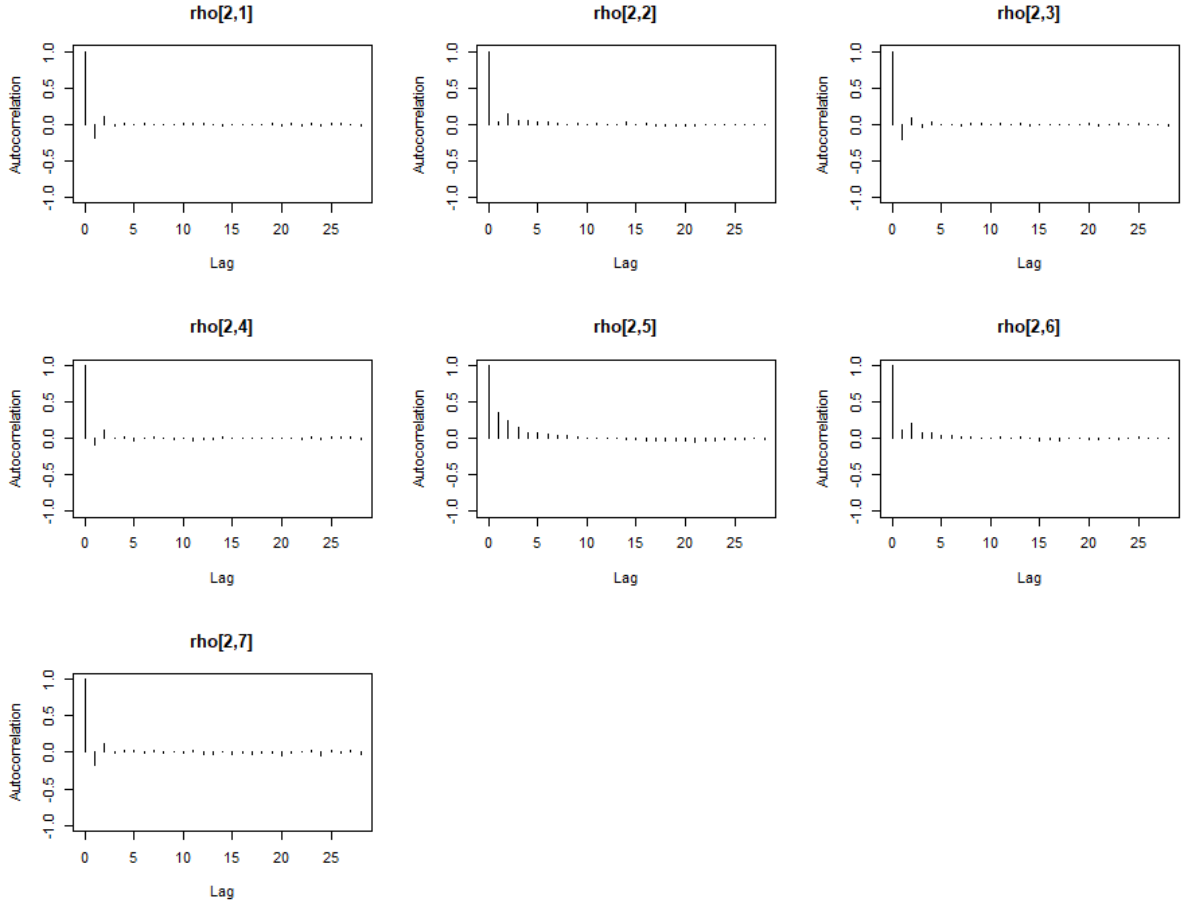


Figura A.9: Gráfico de estatística Gelman-Rubin para os parâmetros $\boldsymbol{\rho_{k,r_k|c=2}}$, de acordo com 5 cadeias geradas pelo método BS para os dados PrEP. 2019-2021

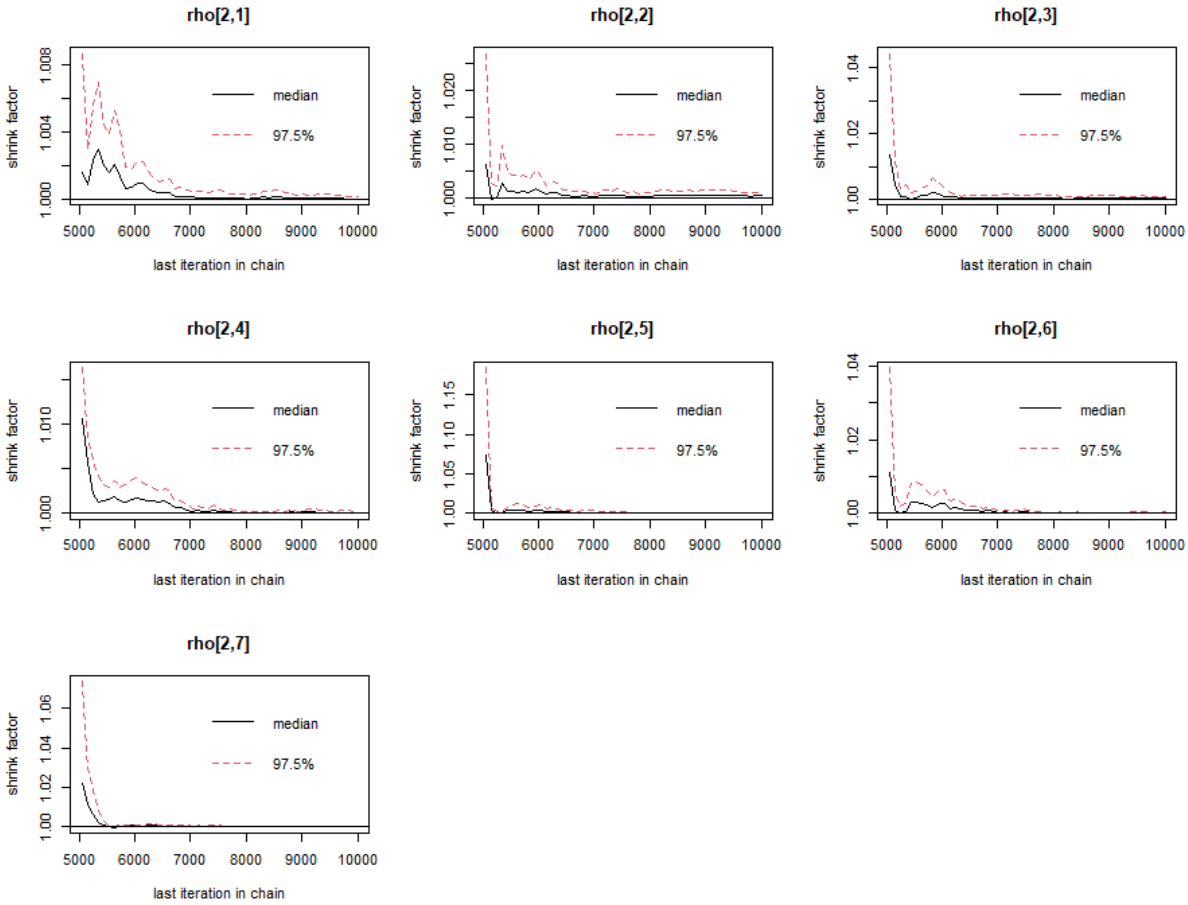


Figura A.10: *Traceplots* e curvas de densidade para os parâmetros ι e $\boldsymbol{\omega}$, de acordo com 5 cadeias geradas pelo método BS para os dados PrEP. 2019-2021

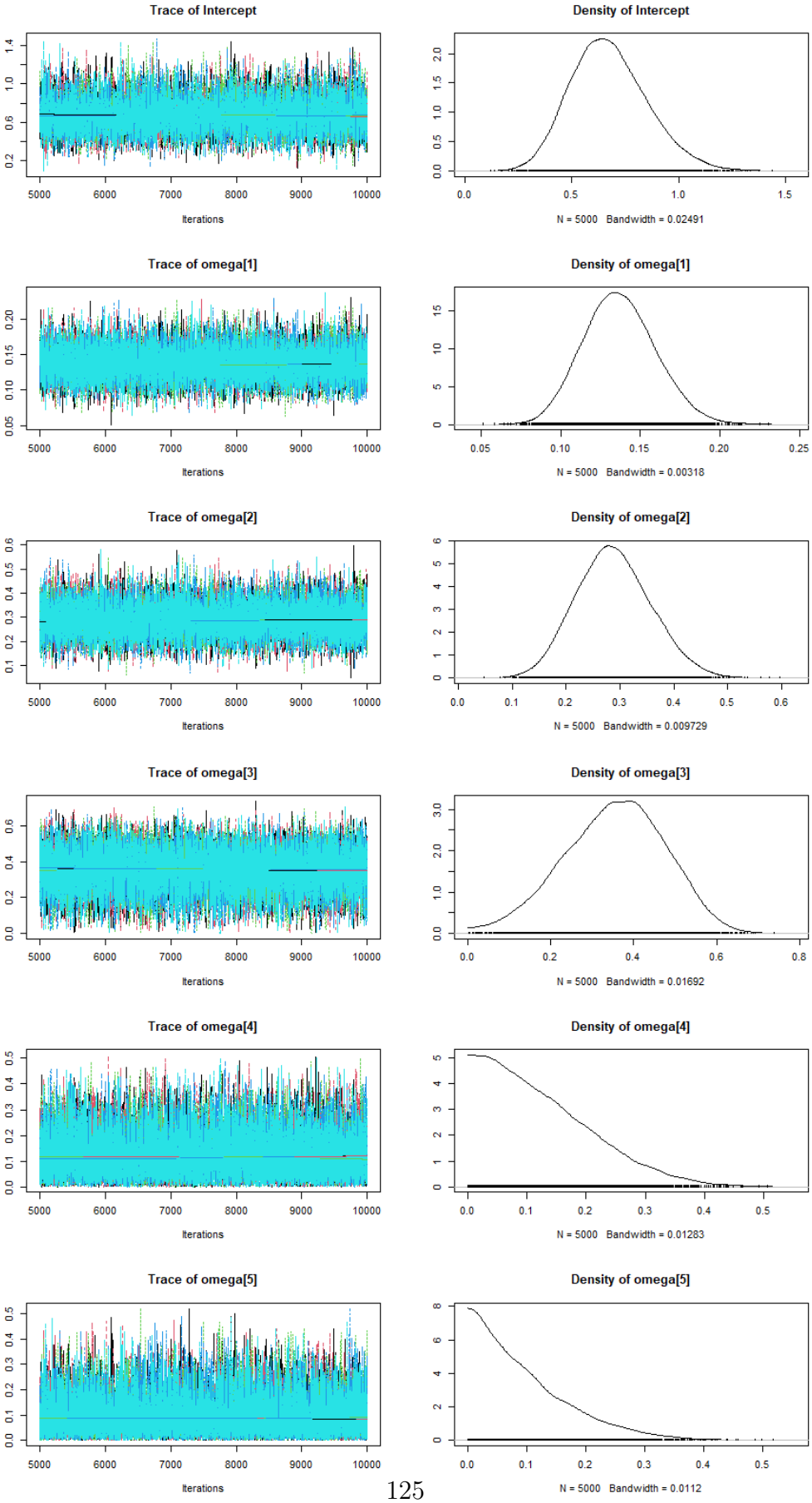


Figura A.11: Autocorrelação para valores amostrados dos parâmetros ι e $\boldsymbol{\omega}$, de acordo com uma cadeia, gerada pelo método BS para os dados PrEP. 2019-2021

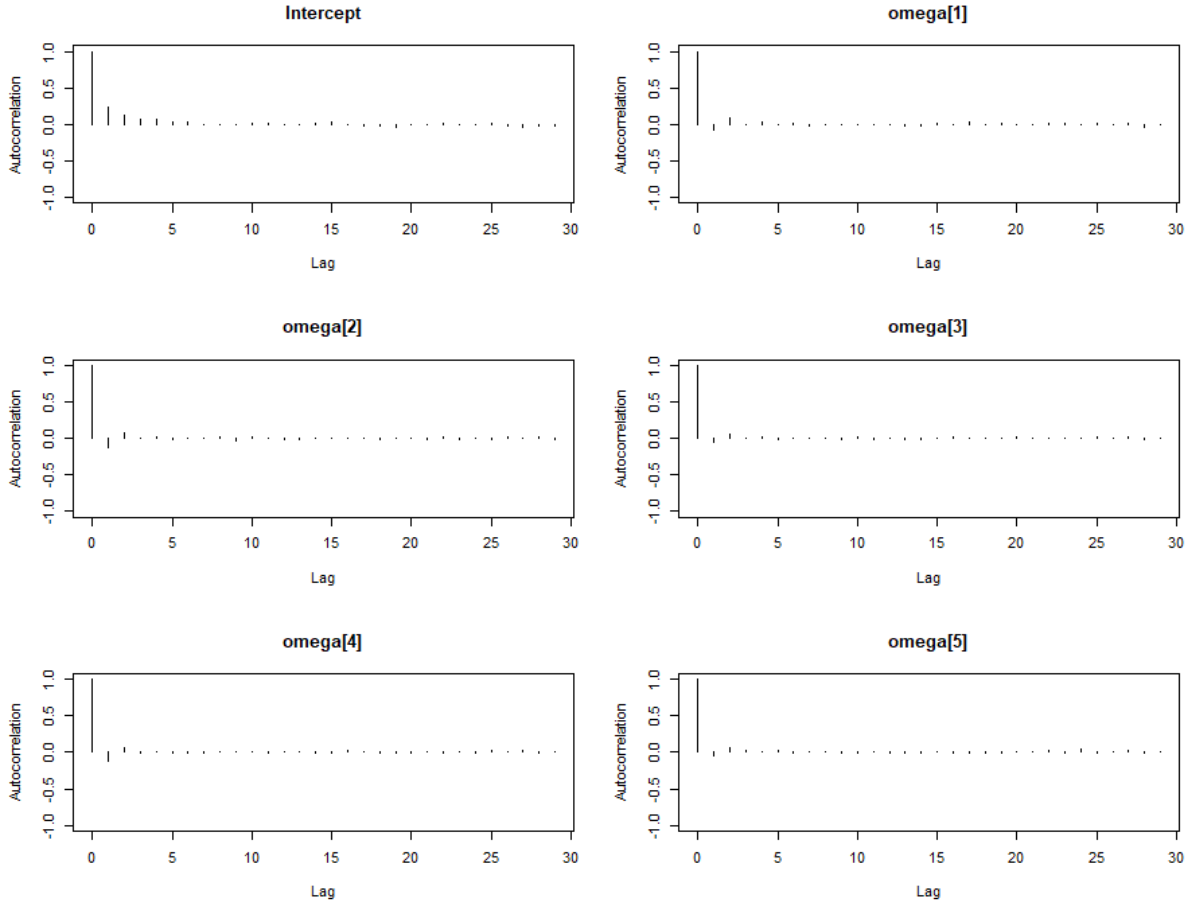
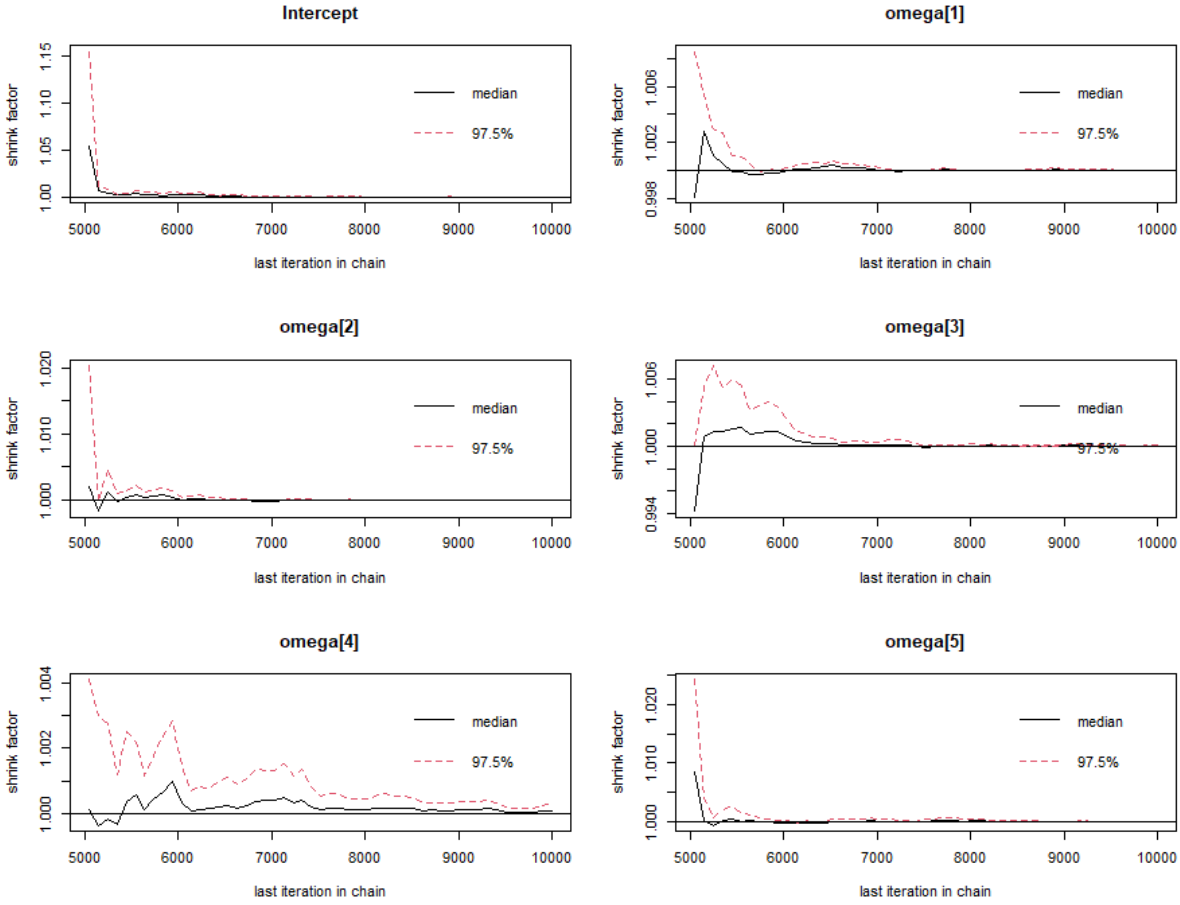


Figura A.12: Gráfico de estatística Gelman-Rubin para os parâmetros ι e $\boldsymbol{\omega}$, de acordo com 5 cadeias geradas pelo método BS para os dados PrEP. 2019-2021



Apêndice B

Investigação sobre valores de parâmetros conforme entropia e percentual de censura

B.1 Valores de λ associados aos percentuais médios de censura de 10 e 30%

Esta seção foi elaborada para ilustrar o procedimento de investigação acerca dos valores para o parâmetro λ necessários para que se tenham amostras com percentuais médios de censura de 10 e 30%, respectivamente. Essas amostras de sobrevivência têm tempos de falha simulados através da metodologia explicitada em Bender, Augustin e Blettner (2005), que apresenta definição necessária para gerar tempos de falha através das distribuições Exponencial, Weibull e Gompertz.

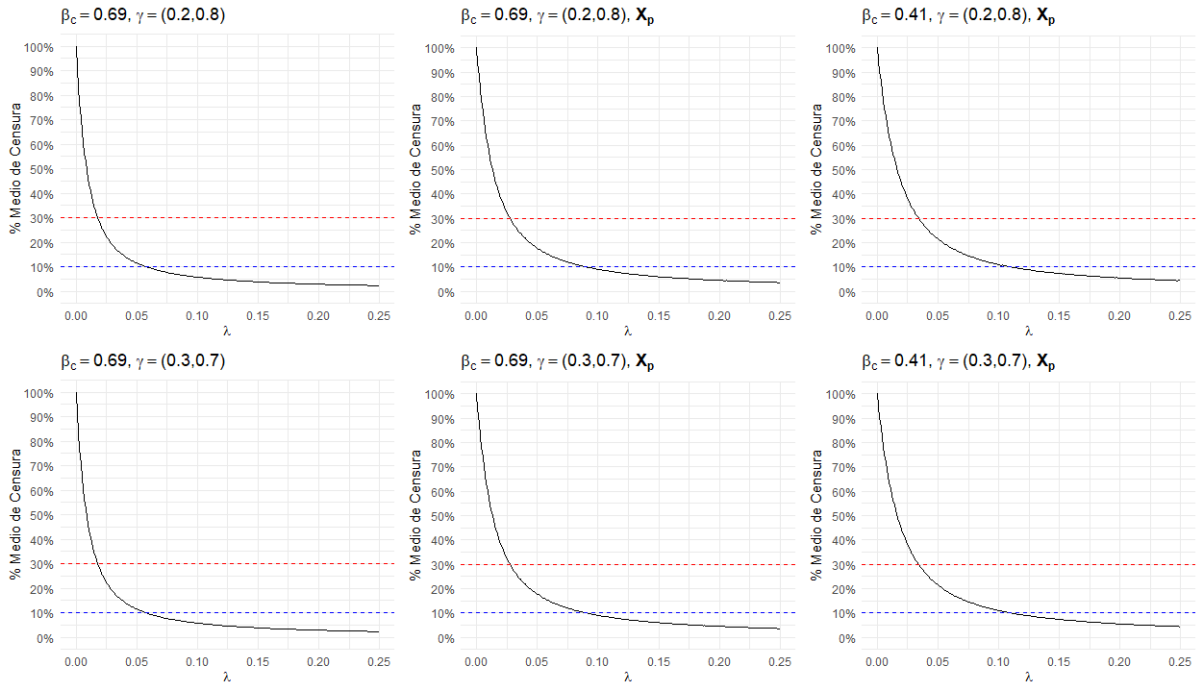
Para simular as amostras de sobrevivência para esta análise foi considerada censura â direita, com mecanismo de censura aleatória. Desta forma são gerados dois vetores de tamanho N contendo tempos de falha e de censura para cada indivíduo, e caso o tempo de falha T_i seja menor que o tempo de censura T_i^* então a falha ocorre, caso contrário o indivíduo teve seu tempo censurado. Este tipo de censura foi escolhido pois é mais próximo do comportamento real dos dados em estudos com indivíduos acompanhados ao longo do tempo e que estão sujeitos a censura (neste caso não informativa) antes do tempo final do estudo T_{max}.

Para esta investigação foram utilizados os seguintes parâmetros de configuração: $N = 1000$, tempo máximo $T_{max} = 104$. A variável binária observada $\mathcal{C}$ foi gerada para representar a variável latente ζ, e desta forma para corresponder aos cenários em que a entropia fosse alta $\mathcal{C} \sim$ Bernoulli(0.2) e para entropia baixa $\mathcal{C} \sim$ Bernoulli(0.3). Os valores para

efeito da variável $\mathcal{C}$ investigados foram $\beta = 0.69$ ou 0.41. Em caso de presença de preditores observados $\mathbf{X}$ considerou-se: $X_1 \sim N(0,1)$, $X_2 \sim \text{Bernoulli}(0.7)$, com coeficientes, respectivamente, iguais a $\varphi_1 = 0.05$ e $\varphi_2 = -0.5$. A covariável $X_3 \sim \text{Multinomial}(0.5, 0.2, 0.3)$ foi incluída no modelo utilizando parametrização que considera o primeiro nível como categoria de referência. Assim, $\varphi_3 = 0.27$ e $\varphi_4 = -0.35$ são os parâmetros associados a X_3. Os valores para $\boldsymbol{\lambda}$ investigados são todos os λ_j com $j = 1, \ldots, 241$ valores contidos no intervalo de $[0; 0.25]$, ao realizar incrementos de 0.001 unidades a partir do limite inferior. Desta forma, para cada valor λ_j contido no intervalo de interesse foram simuladas $w = 1, \ldots, 1000$ amostras, sendo armazenados os seus percentuais de observações censuradas, para que ao fim do processo de simulação das w amostras fosse possível obter o percentual de censura médio associado a cada λ_j.

Após calcular o vetor com os percentuais médios de censura, é possível elaborar um gráfico que relaciona os valores de λ_j com seus respectivos percentuais médios de censura. Neste gráfico, foram adicionadas duas linhas horizontais, nas cores vermelha e azul, representando os percentuais médios de censura de 30% e 10%, respectivamente.

Figura B.1: Valores de λ em relação a percentual de censura médio e preditores observados



A partir dos gráficos apresentados na Figura B.1, é possível obter uma estimativa aproximada dos valores de λ_j que resultam nos percentuais médios de censura desejados. Para identificar quais valores estão mais próximos dos percentuais de censura de interesse, dentre os testados, foram escolhidos aqueles λ_j cujo percentual médio de censura apresentava o menor viés absoluto em relação aos valores de 10% e 30%. Os valores de λ_j correspondentes aos percentuais de censura de interesse, de acordo com as amos-

tras simuladas segundo as configurações dos cenários abordados nesta dissertação estão apresentados na Tabela B.1.

Tabela B.1: Valores de λ de acordo com β_c, percentual de censura e preditores observados.

$\boldsymbol{\beta_c}$	**Sem preditores**		**Com preditores**	
	10% Censura	***30% Censura***	***10% Censura***	***30% Censura***
0.69	0.058	0.018	0.09	0.028
0.41	-	-	0.11	0.034

Os valores apresentados na Tabela B.1 são os valores utilizados para simulação das amostras de tempos de falha utilizadas no Capítulo 5, de forma a obter os percentuais de censura abordados. Ao realizar este estudo não foram encontradas diferenças entre os valores de λ de acordo com as probabilidades utilizadas para gerar a variável ς que representa, nesta notação, a variável latente observada gerada a partir de uma distribuição Bernoulli de acordo com as prevalências de classe latente $\boldsymbol{\gamma} = (0.2, 0.8)$ ou $\boldsymbol{\gamma} = (0.3, 0.7)$. A sintaxe necessária para simulação das amostras utilizadas neste estudo, elaboração dos gráficos e detecção dos valores de λ de interesse estão disponíveis no repositório `distal_bayes_survival` do Github, cuja estrutura é explicada no Apêndice E.

B.2 Valores de $\boldsymbol{\rho}$ associados aos tipos de entropia média padronizada Alta e Baixa

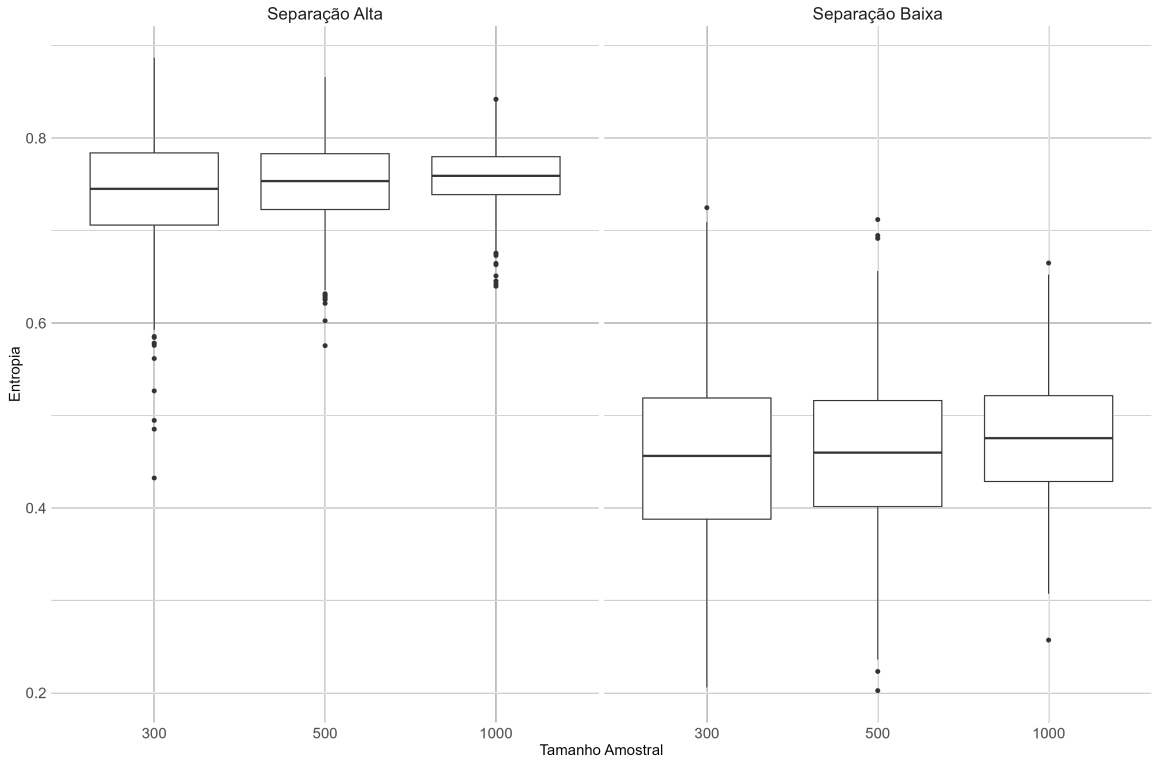
Nesta seção é explicitada a metodologia utilizada para investigação sobre valores para entropia obtidos a partir de modelos de mensuração estimados com amostras geradas de acordo com conjuntos de parâmetros $\boldsymbol{\gamma}$ e $\boldsymbol{\rho}$ que representam as prevalência de classe e as probabilidade condicionais de resposta ao item, respectivamente. Para esta investigação considerou-se indicadores binários e uma variável latente com duas classes.

Considerando tamanhos amostrais com 300, 500 e 1000 indivíduos foram geradas mil amostras de indicadores de classe latente segundo os seguintes conjuntos de parâmetros: **(i)** mais homogêneos com $\boldsymbol{\gamma} = (0.2, 0.8)$, $\boldsymbol{\rho}_1 = (0.8, 0.8, 0.2, 0.2)$ e $\boldsymbol{\rho}_2 = (0.2, 0.2, 0.8, 0.8)$; **(ii)** menos homogêneos com $\boldsymbol{\gamma} = (0.3, 0.7)$, $\boldsymbol{\rho}_1 = (0.8, 0.6, 0.2, 0.4)$ e $\boldsymbol{\rho}_2 = (0.2, 0.7, 0.6, 0.1)$. Após geradas mil amostras de indicadores para cada tamanho amostral, de acordo com os conjuntos de parâmetros abordados visando alta ou baixa separação de classes, são ajustados modelos de mensuração de LCA por meio de modelo definido na Equação (2.32) com implementação via STAN. Após ajuste do modelo, as probabilidades posteriores de classificação, definidas conforme Equação (2.33), foram utilizadas para cálculo da entropia padronizada (Celeux e Soromenho, 1996), que foi

adaptada para o contexto de LCA e definida através da Equação (2.34). Desta forma, foi possível obter valores de entropia correspondente às amostras com diferentes tamanhos amostrais geradas através das combinações de parâmetros relacionadas com alta ou baixa separação de classes. As combinações de parâmetros em particular foram escolhidas baseadas no estudo de Bispo (2019) e após investigação preliminar acerca de quais valores de $\boldsymbol{\rho}$ correspondem à separação de classes alta e baixa, respectivamente, segundo Clark e Muthén (2009).

A Figura B.2 apresenta gráficos do tipo *boxplot* com os valores de entropia padronizada para mil amostras, em cada tamanho amostral, segundo os conjuntos de parâmetros $\boldsymbol{\gamma}$ e $\boldsymbol{\rho}$ abordados para configurações com separações de classe altas ou baixas.

Figura B.2: Valores de entropia segundo tamanho amostral N e separação de classes



A Tabela B.2 apresenta as medidas resumo para os valores de entropia de acordo com os conjuntos de parâmetros do modelo de mensuração investigados.

Tabela B.2: Estatísticas descritivas para entropia segundo tamanho amostral N e separação de classes

Separação	**N**	**Média**	**Mediana**	**Desvio Padrão**	**Mínimo**	**Máximo**
Alta	300	0.74	0.75	0.06	0.43	0.89
	500	0.75	0.75	0.04	0.58	0.87
	1000	0.76	0.76	0.03	0.64	0.84
	Todos	0.75	0.75	0.05	0.43	0.89
Baixa	300	0.45	0.46	0.09	0.21	0.72
	500	0.46	0.46	0.08	0.20	0.71
	1000	0.47	0.48	0.07	0.26	0.66
	Todos	0.46	0.46	0.08	0.20	0.72

Através da Tabela B.2, percebe-se que o conjunto de parâmetros escolhido para investigação resulta em valores médios de entropia padronizada ($\bar{E}$) em torno de 0.75, representando alta entropia, de acordo com a combinação de parâmetros $\boldsymbol{\gamma}$ e $\boldsymbol{\rho}$ que considera maior homogeneidade e separação das classes. Já com o conjunto de parâmetros $\boldsymbol{\gamma}$ e $\boldsymbol{\rho}$ que apresenta menor homogeneidade e separação das classes, obtém-se $\bar{E}$ em torno de 0.45, o que representa baixa entropia. Desta forma foram definidos os parâmetros utilizados para gerar as amostras de indicadores de classe latente, segundo tipo de entropia, utilizadas no Capítulo 5. A sintaxe necessária para gerar as amostras utilizadas neste estudo, elaboração dos gráficos e cálculo da entropia padronizada estão disponíveis no repositório `distal_bayes_survival` do Github, cuja estrutura é explicada no Apêndice E.

Apêndice C

Definições complementares

C.1 *M-splines* e *I-splines*

A base M-spline é avaliada usando o método descrito em Ramsay (1988) e em W. Wang e Yan (2021) de forma que, dada uma sequência de nós simples s_k, a i-ésima função de base *M-spline* de grau k, denotada por $M_{i,k}(x|s_k)$, pode ser considerada como a função de base *B-spline* normalizada que satisfaz:

$$M_{i,k}(x|s_k) = \frac{(k+1)B_{i,k}(x|s_k)}{(\tau_{i+k+1} - \tau_i)} \ .$$

Similar à fórmula recursiva de Cox-de Boor para *B-splines*, a função de base *M-spline* $M_{i,k}(x|s_k)$ pode ser definida recursivamente por (W. Wang e Yan, 2021):

$$M_{i,k}(x|s_k) = \frac{k+1}{k(\tau_{i+k+1} - \tau_i)} \left[(x - \tau_i)M_{i,k-1}(x|s_k) + (\tau_{i+k+1} - x)M_{i+1,k-1}(x|s_k)\right] \ ,$$

com $m \geq 0$ nós internos distintos, $\xi_1, \ldots, \xi_m$, colocados dentro dos nós de fronteira L e U satisfazendo a seguinte condição: $L < \xi_1 < \ldots < \xi_m < U$, em que $\tau_1 = \ldots = \tau_d = L$, $\tau_{d+j} = \xi_j$, $j \in \{1, \ldots, m\}$, e $U = \tau_{d+m+1} = \ldots = \tau_{d+p}$. Os termos k e $d = k + 1$ representam o grau polinomial e a ordem de uma função de base, respectivamente, $p = d + m$ representa os graus de liberdade de uma função de base, e $\sum_{i=1}^{p} M_{i,k}(x|s_k) = 1$, em que $B_{i,k}(x|s_k)$ é a função de base *B-spline* correspondente, $i \in \{1, \ldots, p\}$ (W. Wang e Yan, 2021).

As *M-splines* integradas de τ_1 até x foram denominadas *I-splines* (Ramsay, 1988). Mais especificamente, a i-ésima função de base *I-spline* de grau k, denotada por $I_{i,k}(x|s_k)$ para $i \in \{1, \ldots, p\}$, é definida como $\int_{\tau_1}^{x} M_{i,k}(\tau|s_k)\, d\tau$. Desta forma a função *I-spline* pode ser definida como (Ramsay, 1988; W. Wang e Yan, 2021):

$$I_{i,k}(x|s_k) = \sum_{l=i+1}^{p+1} B_{l,k+1}(x|s_{k+1}) = \sum_{l=i+1}^{p+1} \frac{\tau_{l+k+2} - \tau_l}{k+2} M_{l,k+1}(x|s_{k+1}) \ ,$$

em que τ_{l+k+2} e τ_{l+1} são definidos para a sequência de nós s_{k+1}. As *M-splines* são não negativas para $L \leq x \leq U$. Portanto, as *I-splines* são monotonamente não decrescentes de L para U por definição. Ramsay (1988) propôs o uso de *I-splines* para regressão monotônica. Uma função monotonamente não decrescente (ou não crescente) pode ser ajustada por uma combinação linear de *I-splines* e um termo adicional de intercepto, sendo que a monotonicidade é garantida ao restringir os coeficientes das *I-splines* a serem não negativos (ou não positivos).

C.2 Log-verossimilhança para o modelo completo proposto

Considerando-se o modelo completo, com inclusão dos parâmetros provenientes dos submodelos de mensuração e estrutural, tem-se $\boldsymbol{\varrho} = (\iota, \boldsymbol{\rho}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\varphi}, \boldsymbol{\beta})$. Desta forma, a nova função de log-verossimilhança para estimação destes parâmetros pelo modelo de respostas distais em análise de sobrevivência pode ser expressa através da combinação das Equações (4.4) e (4.1):

$$\begin{aligned}
l(\boldsymbol{\varrho}) &= \sum_{i=1}^{N} \log \left[P(\mathbf{Y}_i = \mathbf{y}_i, \boldsymbol{\xi}_i = \boldsymbol{\varsigma}_i | \boldsymbol{\gamma}, \boldsymbol{\rho}) \cdot P(u_i, \upsilon_i | \ddot{\mathbf{x}}_i, \boldsymbol{\zeta_i}) \right] \\
&= \sum_{i=1}^{N} \log \left[\prod_{c=1}^{C} \left[\gamma_c \prod_{k=1}^{K} \prod_{r_k=1}^{R_k} \rho_{k,r_k|c}^{I(y_{ik}=r_k)} \right]^{\varsigma_{ic}} \cdot P(u_i, \upsilon_i | \ddot{\mathbf{x}}_i, \boldsymbol{\zeta_i}) \right] \\
&= \sum_{i=1}^{N} \left\{ \log \left[\prod_{c=1}^{C} \left[\gamma_c \prod_{k=1}^{K} \prod_{r_k=1}^{R_k} \rho_{k,r_k|c}^{I(y_{ik}=r_k)} \right]^{\varsigma_{ic}} \right] + \log \left[P(u_i, \upsilon_i | \ddot{\mathbf{x}}_i, \boldsymbol{\zeta_i}) \right] \right\} \\
&= \sum_{i=1}^{N} \left\{ \left[\sum_{c=1}^{C} \log \left[\gamma_c \prod_{k=1}^{K} \prod_{r_k=1}^{R_k} \rho_{k,r_k|c}^{I(y_{ik}=r_k)} \right]^{\varsigma_{ic}} \right] + \log \left[P(u_i, \upsilon_i | \ddot{\mathbf{x}}_i, \boldsymbol{\zeta_i}) \right] \right\} \\
&= \sum_{i=1}^{N} \left\{ \sum_{c=1}^{C} \varsigma_{ic} \left[\log(\gamma_c) + \log \left(\prod_{k=1}^{K} \prod_{r_k=1}^{R_k} \rho_{k,r_k|c}^{I(y_{ik}=r_k)} \right) \right] + \log \left[P(u_i, \upsilon_i | \ddot{\mathbf{x}}_i, \boldsymbol{\zeta_i}) \right] \right\} \ .
\end{aligned}$$

De acordo com a definição do termo $P(u_i, \upsilon_i|\ddot{\mathbf{x}}_i, \boldsymbol{\zeta_i})$ segundo Equação (4.4), temos o seguinte resultado ao aplicar o logaritmo natural:

$$\log[P(u_i, \upsilon_i|\ddot{\mathbf{x}}_i, \boldsymbol{\zeta_i})] = \upsilon_i \cdot \left[\log\left(\sum_{l=1}^{L} \omega_l M_l(u_i; \boldsymbol{\kappa}, \delta)\right) + (\iota + \ddot{\mathbf{x}}_i\boldsymbol{\varphi} + \zeta_{\mathbf{i}l}\boldsymbol{\beta})\right] - \sum_{l=1}^{L} \omega_l I_l(u_i; \boldsymbol{\kappa}, \delta) \cdot \exp(\iota + \ddot{\mathbf{x}}_i\boldsymbol{\varphi} + \zeta_{\mathbf{i}}\boldsymbol{\beta}) \ .$$

Logo a log-verossimilhança para o modelo completo pode ser escrita como:

$$l(\boldsymbol{\varrho}) = \sum_{i=1}^{N} \left\{ \sum_{c=1}^{C} \varsigma_{ic} \left[\log(\gamma_c) + \log\left(\prod_{k=1}^{K} \prod_{r_k=1}^{R_k} \rho_{k,r_k|c}^{I(y_{ik}=r_k)}\right)\right] + \upsilon_i \cdot \left[\log\left(\sum_{l=1}^{L} \omega_l M_l(u_i; \boldsymbol{\kappa}, \delta)\right) + (\iota + \ddot{\mathbf{x}}_i\boldsymbol{\varphi} + \zeta_{\mathbf{i}}\boldsymbol{\beta})\right] - \sum_{l=1}^{L} \omega_l I_l(u_i; \boldsymbol{\kappa}, \delta) \cdot \exp(\iota + \ddot{\mathbf{x}}_i\boldsymbol{\varphi} + \zeta_{\mathbf{i}}\boldsymbol{\beta}) \right\} \ .$$

Apêndice D

Aspectos computacionais dos estudos de simulação

Este apêndice descreve as características computacionais, os tempos necessários para a realização dos estudos de simulação, além da frequência (absoluta e relativa) de problemas de convergência por cenário, referentes ao estudo apresentado no Capítulo 5.

D.1 Tempos de estimação por cenário

Foram computados os tempos necessários para a estimação dos mil modelos iniciais de acordo com os métodos BS e BSM, baseados nas mil amostras geradas para os tamanhos amostrais 300, 500 e 1000, em cada um dos 12 cenários abordados. Nos cenários 1 a 4, definidos na Tabela 5.1 (Seção 5.2), o efeito associado à variável latente apresenta magnitude alta ($\beta_c = 0.69$) e não há preditores observados. Já nos cenários 5 a 12, que incluem preditores observados $\ddot{\mathbf{X}}$, a magnitude do efeito associado à variável latente varia, sendo $\beta_c = 0.69$ para os cenários 5 a 8 e $\beta_c = 0.41$ para os cenários 9 a 12.

Os tempos de estimação estão apresentados na Tabela D.1, de acordo com cenário, método de estimação e tamanho amostral.

Tabela D.1: Tempos de execução para mil modelos nos cenários analisados de acordo com tamanho amostral e método de estimação.

Entropia	Censura	Cenário	*Método BS*				*Método BSM*			
			N=300	N=500	N=1000	Total	N=300	N=500	N=1000	Total
Alta	10%	1	57M : 51S	1H : 45M : 8S	3H : 53M : 11S	6H : 36M : 10S	38M : 19S	1H : 1M : 36S	2H : 5M : 3S	3H : 44M : 58S
	30%	2	55M : 24S	1H : 30M : 38S	3H : 10M : 37S	5H : 36M : 39S	36M : 17S	57M : 24S	1H : 52M : 3S	3H : 25M : 44S
Baixa	10%	3	3H : 22M : 34S	4H : 3M : 18S	9H : 18M : 35S	16H : 44M : 27S	1H : 42M : 15S	2H : 29M : 15S	5H : 50M : 6S	10H : 1M : 36S
	30%	4	5H : 53M : 5S	5H : 42M : 14S	8H : 1M : 45S	19H : 37M : 4S	1H : 53M : 50S	2H : 57M : 9S	5H : 42M : 20S	10H : 33M : 19S
Alta	10%	5	1H : 6M : 1S	1H : 59M : 3S	4H : 32M : 2S	7H : 37M : 6S	41M : 28S	1H : 7M : 49S	2H : 16M : 30S	4H : 5M : 48S
	30%	6	1H : 3M : 25S	1H : 46M : 45S	3H : 50M : 4S	6H : 40M : 14S	42M : 56S	1H : 6M : 19S	2H : 0M : 28S	3H : 49M : 43S
Baixa	10%	7	5H : 59M : 30S	6H : 24M : 16S	13H : 37M : 1S	1D : 2H : 0M : 46S	1H : 30M : 17S	2H : 25M : 2S	5H : 38M : 46S	9H : 34M : 5S
	30%	8	12H : 45M : 46S	8H : 47M : 37S	13H : 29M : 46S	1D : 11H : 3M : 9S	1H : 57M : 57S	3H : 8M : 40S	6H : 13M : 56S	11H : 20M : 33S
Alta	10%	9	54M : 2S	1H : 58M : 9S	4H : 23M : 16S	7H : 15M : 27S	1H : 1M : 12S	1H : 39M : 29S	3H : 13M : 44S	5H : 54M : 24S
	30%	10	1H : 20M : 18S	2H : 9M : 16S	4H : 37M : 60S	8H : 7M : 35S	1H : 0M : 8S	1H : 32M : 9S	2H : 52M : 23S	5H : 24M : 40S
Baixa	10%	11	10H : 49M : 29S	7H : 8M : 39S	13H : 59M : 19S	1D : 7H : 57M : 26S	2H : 3M : 6S	3H : 10M : 58S	6H : 53M : 2S	12H : 7M : 5S
	30%	12	13H : 49M : 30S	12H : 29M : 28S	14H : 18M : 10S	1D : 16H : 37M : 7S	1H : 32M : 51S	2H : 27M : 25S	4H : 53M : 49S	8H : 54M : 5S

De acordo com a Tabela D.1 nota-se que os tempos de execução para mil modelos, utilizando técnica de computação paralela através dos pacotes `future` e `doFuture`, tendem a aumentar de acordo com aumentos no tamanho amostral, percentual de censura, complexidade do modelo (presença de preditores observados $\ddot{\mathbf{X}}$) e diminuição da entropia. Este comportamento pode ser observado analisando-se por tamanho amostral ou por tempo total. Os modelos estimados via método BS apresentaram tempos maiores para execução se comparados com o método BSM, mesmo o método BSM realizando ajustes separados para os submodelos de mensuração e estrutural. Além disto, em relação ao método BS, o acréscimo do tempo de estimação é mais notável em cenários com baixa entropia, chegando a ser de 3 a 12 vezes maior se comparados com os tempos nos cenários com alta entropia.

Para estimação dos modelos foram utilizados 20 núcleos de CPU simultâneos. O computador utilizado conta com processador AMD MILAN 7713 DP/UP 64C/128T operando a uma frequência de 2.0 GHz. Para estimação dos modelos em cada um dos cenários abordados foram utilizados 256 *gigabytes* de memória RAM DDR4 operando à uma frequência de 3200 MHz. O sistema operacional utilizado foi o Linux CentOS.

D.2 Quantidade e percentual de problemas de convergência por cenário

A Tabela D.2 apresenta as frequências (absoluta e relativa) de problemas de convergência detectados por cenário em cada um dos tamanhos amostrais analisados. Esses valores referem-se ao método de estimação simultânea BS. O número de problemas de convergência no método BSM é similar ao do método BS, pois caso ocorra um problema de convergência na execução de um modelo associado a um conjunto de dados via método BS, e o mesmo problema não seja identificado no método BSM, o conjunto de dados é

substituído por outro, gerado com as mesmas configurações de parâmetros do anterior, e a estimação é realizada novamente para ambos os métodos.

Tabela D.2: Frequências (absolutas e relativas) de problemas de convergência na execução de mil modelos nos cenários analisados, de acordo com o tamanho amostral e o método de estimação BS.

Entropia	**Censura**	**Cenário**	***Método BS***		
			N=300 (%)	**N=500 (%)**	**N=1000 (%)**
Alta	10%	1	2 (0.2)	2 (0.2)	0 (0)
	30%	2	3 (0.3)	0 (0)	0 (0)
Baixa	10%	3	81 (8.1)	6 (0.6)	5 (0.5)
	30%	4	184 (18.4)	39 (3.9)	8 (0.8)
Alta	10%	5	0 (0)	0 (0)	0 (0)
	30%	6	1 (0.1)	1 (0.1)	2 (0.2)
Baixa	10%	7	75 (7.5)	6 (0.6)	4 (0.4)
	30%	8	212 (21.2)	25 (2.5)	4 (0.4)
Alta	10%	9	0 (0)	2 (0.2)	1 (0.1)
	30%	10	2 (0.2)	2 (0.2)	0 (0)
Baixa	10%	11	215 (21.5)	16 (1.6)	7 (0.7)
	30%	12	272 (27.2)	48 (4.8)	22 (2.2)

De acordo com a Tabela D.2, o número de problemas de convergência aumenta com reduções do tamanho amostral e entropia, e com o aumento dos percentuais de censura. Nos cenários com alta entropia, o maior número de problemas de convergência detectado foi 3 (cenário 2), representando 0,03% dos mil modelos iniciais estimados por cenário. Por outro lado, nos cenários com baixa entropia, observa-se um aumento acentuado na frequência de problemas de convergência, com os piores resultados ocorrendo no menor tamanho amostral, maior percentual de censura e presença de preditores observados $\ddot{\mathbf{X}}$ de ζ, totalizando 272 problemas (27,2%) detectados.

Apêndice E

Documentação e estrutura do repositório Github

Neste Apêndice é detalhada estrutura do repositório Github `distal_bayes_survival` em que estão armazenadas as sintaxes computacionais necessárias para estimação do modelo proposto segundo os métodos BS e BSM, e simulação das amostras utilizadas. Esse diretório é subdividido em quatro pastas principais:

- `method_bs`: Contém sintaxe necessária para implementação do modelo proposto no Capítulo 4 segundo método BS;
- `method_bsm`: Contém sintaxe necessária para implementação do modelo proposto no Capítulo 4 segundo método BSM;
- `research_lambda_rho`: Contém sintaxe necessária para investigação dos valores de $\boldsymbol{\rho}$ e λ necessários de acordo com entropia e percentual de censura de interesse;
- `generate_samples`: Contém sintaxe necessária para gerar amostras utilizadas nos estudos de simulação apresentados no Capítulo 5 de acordo com os cenários abordados.

Nas subpastas `method_bs` e `method_bsm` constam, além da sintaxe em linguagem STAN para implementação do modelo, arquivos em extensão `.R` contendo exemplos de como realizar o ajuste do modelo utilizando interface `Rstan` via *software* R. O repositório conta ainda com arquivos `README.md` contendo orientações mais detalhadas de acordo com os arquivos presentes cada uma das subpastas. O acesso ao repositório `distal_bayes_survival` pode ser realizado após solicitação e aprovação de acesso.